

Proyecto 1

Luis Pedro Lira (23669) Fernando Rocha (23501)
Juan Francisco Martínez (23617) Luis Gilberto González (23353)
Joel Antonio Jaquez (23369)

2026-02-02

Introducción

La violencia intrafamiliar constituye un problema social de alta relevancia debido a su impacto en la estabilidad emocional, social y económica de los hogares. Por otro lado, el divorcio representa una consecuencia legal de conflictos persistentes dentro del núcleo familiar, los cuales pueden estar relacionados con dinámicas de violencia y deterioro de la convivencia.

En este proyecto se analiza la posible relación entre los divorcios y la violencia intrafamiliar utilizando datos oficiales correspondientes al año 2022. A través de técnicas de análisis exploratorio de datos y minería de datos, se busca identificar patrones, comportamientos y relaciones entre ambas problemáticas, con el fin de obtener una comprensión más profunda de su interacción a nivel agregado.

El análisis se centra en una exploración descriptiva inicial de los conjuntos de datos, la caracterización de sus variables y la evaluación de posibles relaciones estadísticas entre los indicadores de divorcio y violencia intrafamiliar.

Objetivos

Objetivo General Analizar la relación entre los divorcios y la violencia intrafamiliar mediante técnicas de análisis exploratorio y minería de datos, utilizando información correspondiente al año 2022.

Objetivos específicos - Describir las características principales de los conjuntos de datos de divorcios y violencia intrafamiliar. - Identificar y clasificar las variables numéricas y categóricas presentes en cada conjunto de datos. - Realizar un análisis exploratorio de las variables relevantes mediante medidas descriptivas y representaciones gráficas. - Evaluar posibles relaciones entre indicadores de divorcio y violencia intrafamiliar.

Avances

Descripción general de los datos

Para el desarrollo del presente proyecto se utilizaron dos conjuntos de datos correspondientes al año 2022. El primero contiene información relacionada con los procesos de divorcio, mientras que el segundo recopila registros asociados a casos de violencia intrafamiliar. Ambos conjuntos de datos fueron obtenidos a partir de fuentes oficiales y procesados previamente para su uso en el análisis.

Descripción de las variables: Violencia intrafamiliar

El conjunto de datos de violencia intrafamiliar contiene información detallada sobre las características sociodemográficas de las víctimas y agresores, así como aspectos relacionados con el hecho de violencia reportado. Las variables incluidas permiten analizar dimensiones como edad, sexo, escolaridad, ocupación, estado civil y relación entre víctima y agresor, entre otras.

La mayoría de las variables en el conjunto de datos corresponden a codificaciones numéricas que representan categorías, mientras que otras variables como la edad y las fechas corresponden a valores numéricos continuos o discretos

Descripción de las variables: Divorcios

El conjunto de datos de divorcios contiene información asociada a los procesos de disolución del vínculo matrimonial, incluyendo características temporales, geográficas y sociodemográficas de las personas involucradas. Las variables disponibles permiten analizar aspectos como la edad, escolaridad, ocupación y nacionalidad de los cónyuges, así como el lugar y fecha de ocurrencia del divorcio.

Exploración de variables numéricas

En esta sección se analizan las principales variables numéricas de los conjuntos de datos de violencia intrafamiliar y divorcios, utilizando medidas de tendencia central, medidas de orden y representaciones gráficas.

```
##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.
##      1.0   24.0   31.0    34.2   40.0    99.0
```

La edad de la víctima presenta un rango amplio, desde 1 hasta 98 años, lo cual evidencia que la violencia intrafamiliar afecta a personas de distintas etapas de la vida. La mediana de edad es de 31 años, mientras que la media es de aproximadamente 33 años, lo que indica una ligera asimetría positiva en la distribución. El 50% de las víctimas se concentra entre los 24 y 39 años de edad, sugiriendo que los adultos jóvenes representan el grupo más afectado. Se identificaron 5,046 registros sin información de edad, los cuales serán considerados como valores faltantes en el análisis.

```
##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.    NA's
##      0.0   1.0   2.0    20.5    5.0    99.0    1694
```

El número total de hijos de la víctima presenta una mediana de 2 hijos, lo que indica que la mayoría de los casos corresponde a personas con familias pequeñas. La media es cercana a la mediana, lo cual sugiere una distribución relativamente equilibrada. El 75% de las víctimas tiene tres hijos o menos, mientras que los valores máximos observados corresponden a casos aislados de familias numerosas. Se identificó una proporción considerable de valores faltantes, los cuales fueron tratados adecuadamente durante el análisis exploratorio.

```
##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.
##     15.0   31.0   54.0   501.5   999.0   999.0
```

La edad de la mujer en los procesos de divorcio presenta una mediana de 31 años y una media de aproximadamente 33 años, lo que indica que los divorcios se concentran principalmente en edades adultas jóvenes. El 50% de los casos se encuentra entre los 26 y 38 años, mientras que los valores máximos corresponden a casos aislados de edades avanzadas. La distribución muestra una ligera asimetría positiva y se identifican valores faltantes que fueron considerados adecuadamente durante el análisis.

```
##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.  
##    15.0    34.0    61.0   504.2   999.0   999.0
```

La edad del hombre en los procesos de divorcio presenta una mediana de 34 años y una media de aproximadamente 36 años, lo que sugiere que los hombres tienden a ser ligeramente mayores que las mujeres al momento del divorcio. El rango intercuartílico se ubica entre los 29 y 41 años, concentrando la mayoría de los casos en edades adultas. La distribución presenta una asimetría positiva y algunos valores extremos correspondientes a edades avanzadas.

Metodología y explicación de procedimientos

Tratamiento de valores faltantes y códigos especiales

Antes de realizar el análisis estadístico, se revisaron los conjuntos de datos para identificar valores faltantes y códigos especiales utilizados para representar información no especificada. En ambos conjuntos se encontraron valores como 99 y 999, que no representan datos reales sino ausencia de información.

Estos valores fueron transformados en NA con el objetivo de evitar distorsiones en las medidas de tendencia central y en los análisis posteriores. Esta decisión metodológica permitió obtener estadísticas más representativas del comportamiento real de las variables estudiadas.

Análisis descriptivo de variables numéricas

Para las variables cuantitativas seleccionadas se calcularon medidas de tendencia central (media y mediana) y medidas de orden (cuartiles). Estas estadísticas permitieron identificar la concentración de los datos y evaluar la dispersión de cada variable.

En el caso de la edad de las víctimas de violencia intrafamiliar, se observó que la mediana se concentra en adultos jóvenes, lo cual indica que este grupo etario representa una proporción importante de los casos registrados. Asimismo, la ligera diferencia entre media y mediana sugiere una asimetría moderada en la distribución.

En los procesos de divorcio, se identificó que los hombres tienden a ser ligeramente mayores que las mujeres al momento del divorcio. Este patrón consistente refuerza la coherencia interna del conjunto de datos.

Interpretación de la distribución y valores extremos

El análisis exploratorio permitió observar que las variables numéricas no siguen una distribución perfectamente normal, sino que presentan asimetrías moderadas, lo cual es común en fenómenos sociales reales.

Los valores extremos fueron examinados para determinar si correspondían a errores de registro o a casos poco frecuentes pero posibles. En general, los valores observados se consideran coherentes dentro del contexto del fenómeno estudiado, por lo que no fueron eliminados arbitrariamente.

Coherencia y consistencia de los datos

Los resultados obtenidos muestran patrones consistentes entre los diferentes análisis realizados. La concentración de casos en adultos jóvenes tanto en violencia intrafamiliar como en divorcios sugiere dinámicas sociales relevantes que merecen un análisis más profundo.

Es importante destacar que hasta este punto del estudio se han identificado asociaciones descriptivas, sin establecer relaciones causales. Los análisis posteriores permitirán profundizar en la comprensión de estas dinámicas. _____

Exploración de variables categóricas

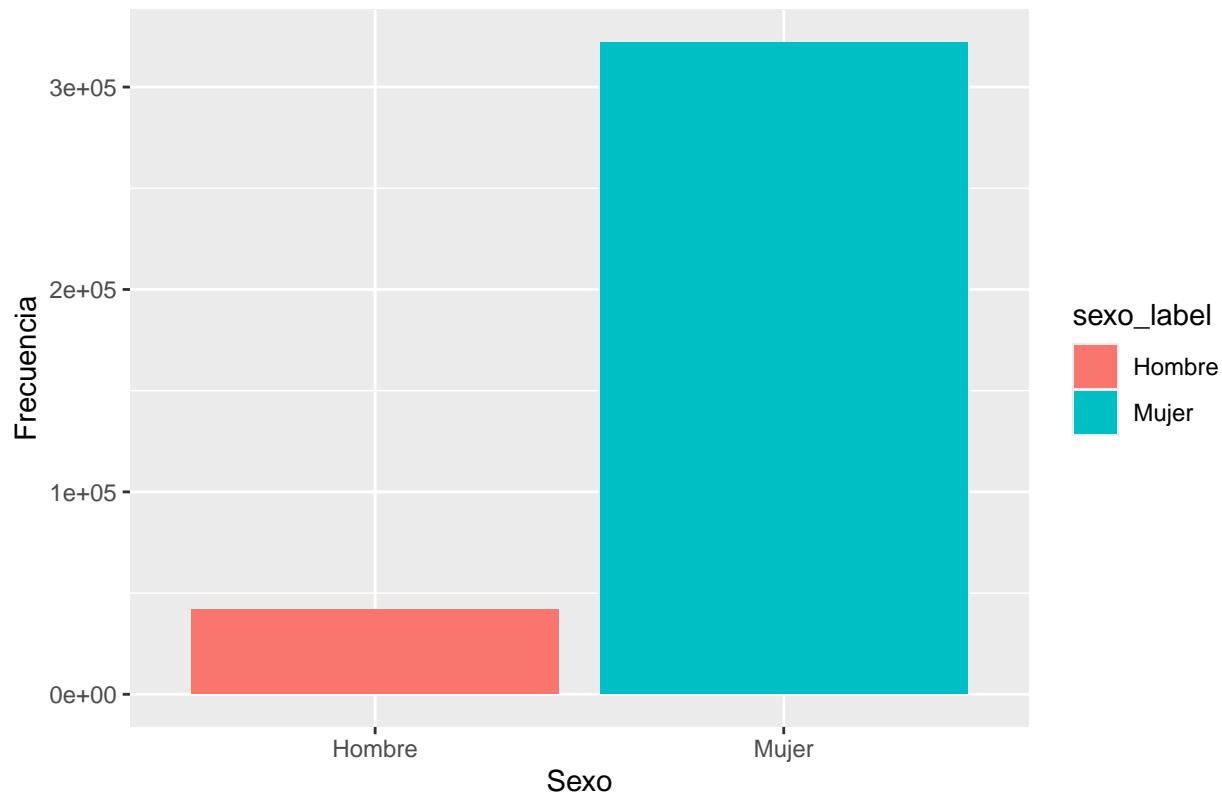
En esta sección se analizan las principales variables categóricas del conjunto de datos de violencia intrafamiliar. El objetivo es identificar la distribución de los casos según características sociodemográficas y relacionales, utilizando tablas de frecuencia y representaciones gráficas que permitan una mejor comprensión del fenómeno estudiado.

Sexo de la víctima

Se analiza la distribución de los casos de violencia intrafamiliar según el sexo de la víctima, con el fin de identificar posibles diferencias en la afectación por género.

```
## # A tibble: 2 x 3
##   sexo_label     n porcentaje
##   <chr>       <int>     <dbl>
## 1 Hombre        41796     11.5
## 2 Mujer        322092    88.5
```

Distribucion del sexo de las victimas



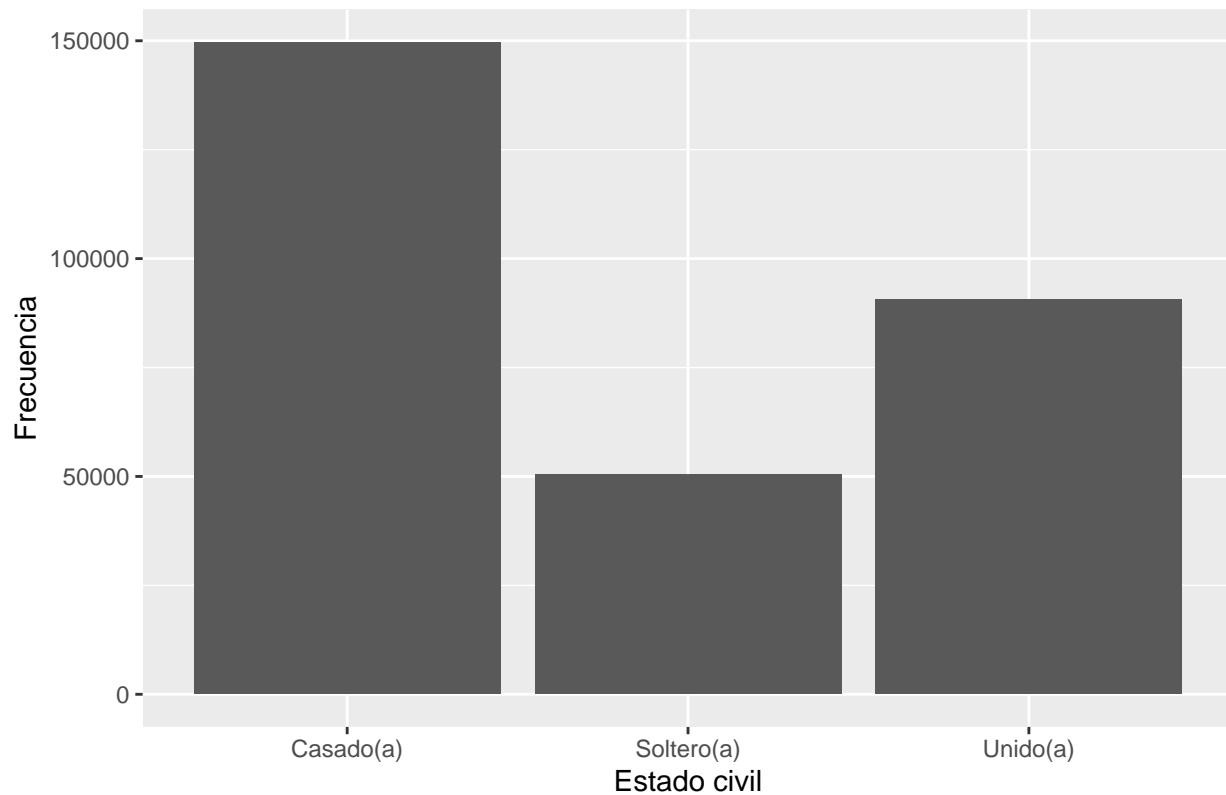
La distribución evidencia que la violencia intrafamiliar afecta de manera diferenciada según el sexo de la víctima, mostrando una concentración mayor en uno de los grupos, lo cual sugiere la presencia de desigualdades estructurales en el contexto familiar.

Estado civil de la víctima

El estado civil de la víctima permite analizar el contexto relacional en el que ocurren los hechos de violencia intrafamiliar.

```
## # A tibble: 3 x 3
##   est_civ_label     n  porcentaje
##   <chr>       <int>      <dbl>
## 1 Casado(a)    149639      51.5
## 2 Soltero(a)    50472      17.4
## 3 Unido(a)     90564      31.2
```

Distribucion del estado civil de las victimas



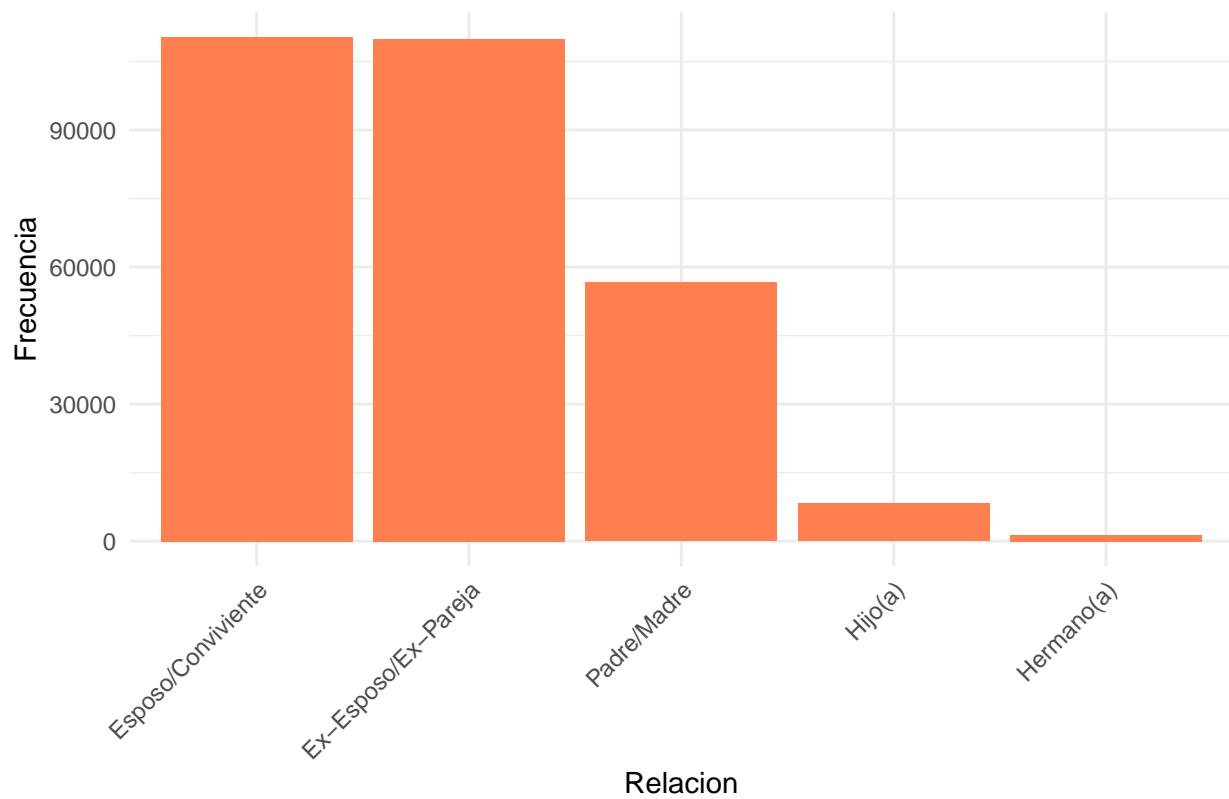
Los resultados muestran que los casos de violencia intrafamiliar se concentran principalmente en ciertos estados civiles, lo cual podría estar asociado a dinámicas de convivencia y conflictos dentro del núcleo familiar.

Relación entre la víctima y el agresor

Se examina la relación existente entre la víctima y el agresor para identificar los vínculos más frecuentes en los casos reportados de violencia intrafamiliar.

```
## # A tibble: 5 x 3
##   relacion_label      n  porcentaje
##   <chr>        <int>     <dbl>
## 1 Esposo/Conviviente 110364     38.5
## 2 Ex-Esposo/Ex-Pareja 109926     38.4
## 3 Hermano(a)         1367      0.477
## 4 Hijo(a)            8315      2.90
## 5 Padre/Madre        56559     19.7
```

Relacion entre la victim a y el agresor



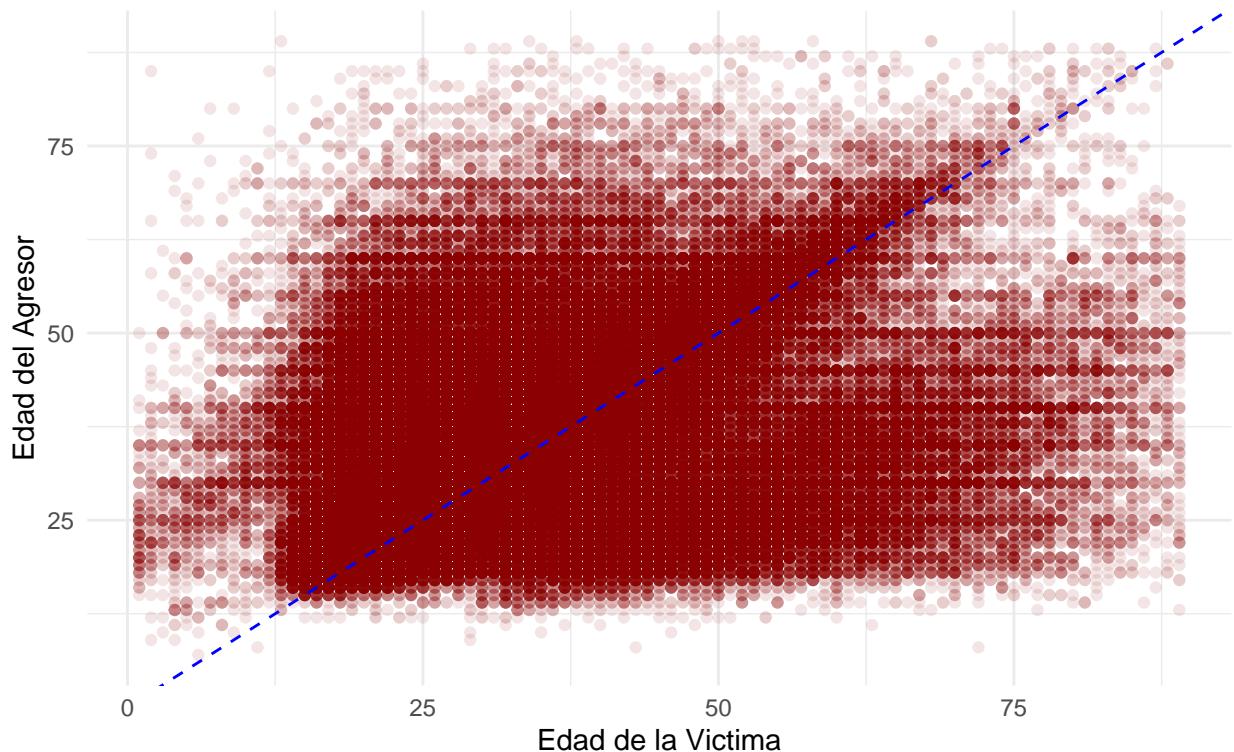
El análisis evidencia que la mayoría de los casos de violencia intrafamiliar ocurre entre personas con vínculos cercanos, lo cual refuerza la naturaleza intrafamiliar del fenómeno y resalta la importancia de analizar las relaciones personales dentro del hogar.

Relaciones entre las variables.

Relación: Diferencia de edad entre Agresor y Víctima

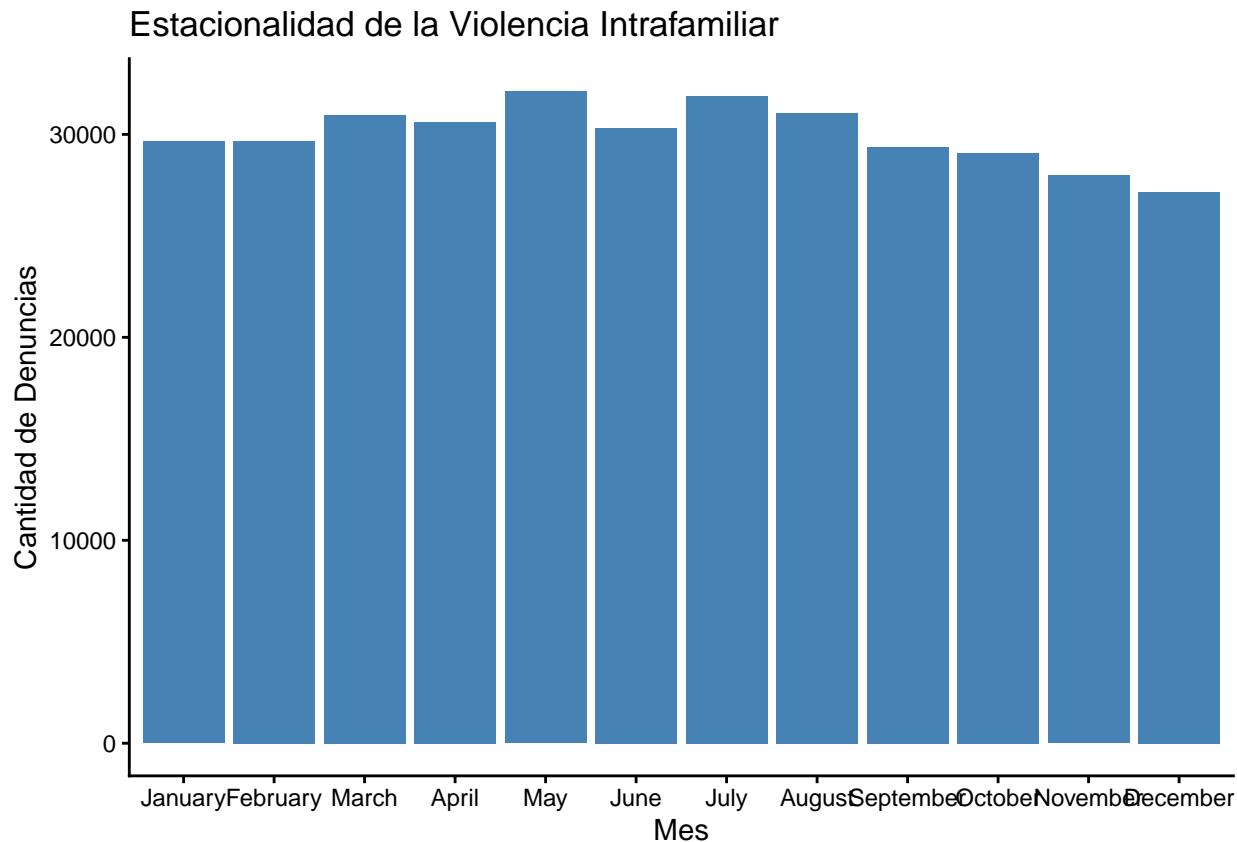
Relacion de Edades: Victima vs. Agresor

La linea azul indica la misma edad. Puntos arriba = Agresor mayor.



El gráfico de dispersión analiza la brecha generacional existente entre la víctima y el agresor en los casos de violencia intrafamiliar. La línea azul discontinua representa la igualdad de edades; por lo tanto, la notable concentración de puntos por encima de esta referencia evidencia que, en la mayoría de los casos reportados, el agresor tiende a ser mayor que la víctima, lo cual sugiere una posible asimetría de poder vinculada a la edad. Adicionalmente, se observa una alta densidad de incidentes (la “nube” más oscura) donde ambas partes oscilan entre los 20 y 40 años, identificando a los adultos jóvenes como el grupo demográfico más vulnerable y activo en esta problemática.

Relación: Estacionalidad de la Violencia (Mes del hecho)

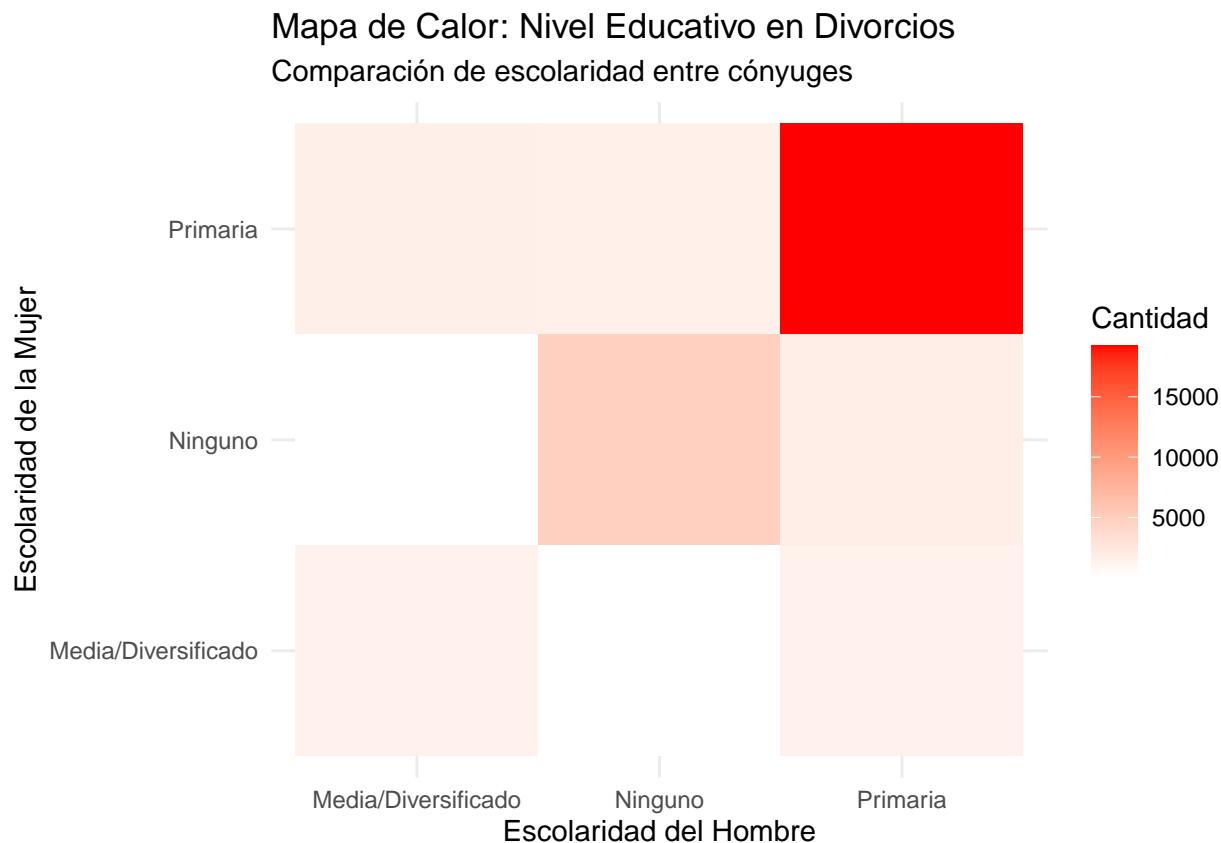


```
## <theme> List of 1
## $ axis.text.x: <ggplot2::element_text>
##   ..@ family      : NULL
##   ..@ face        : NULL
##   ..@ italic       : chr NA
##   ..@ fontweight  : num NA
##   ..@ fontwidth   : num NA
##   ..@ colour      : NULL
##   ..@ size         : NULL
##   ..@ hjust        : num 1
##   ..@ vjust        : NULL
##   ..@ angle        : num 45
##   ..@ lineheight   : NULL
##   ..@ margin       : NULL
##   ..@ debug        : NULL
##   ..@ inherit.blank: logi FALSE
##   @ complete: logi FALSE
##   @ validate: logi TRUE
```

El análisis de estacionalidad revela que la incidencia de violencia intrafamiliar mantiene un comportamiento relativamente constante y sostenido durante gran parte del año, sin caídas drásticas. Se observan ligeros repuntes en los meses de marzo, mayo y julio, los cuales concentran los volúmenes más altos de denuncias (superando los 30,000 casos mensuales). Por otro lado, se percibe una leve tendencia decreciente hacia el

último trimestre (octubre a diciembre). Cabe destacar la presencia de una barra etiquetada como “99”, la cual corresponde a registros donde el mes de ocurrencia no fue especificado o es desconocido, evidenciando la necesidad de considerar estos valores nulos durante la limpieza de datos para evitar sesgos en el análisis temporal.

Relación: Nivel Educativo de la Pareja (Homogamia)



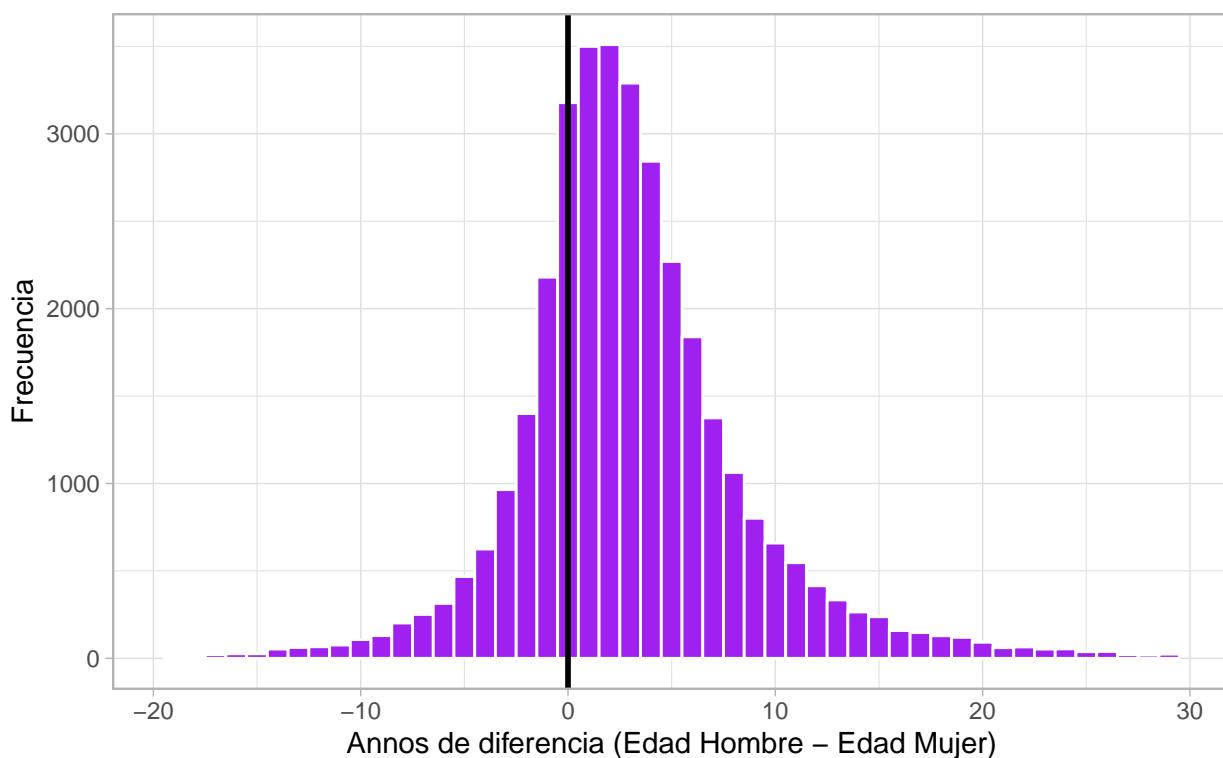
```
## <theme> List of 1
## $ axis.text.x: <ggplot2::element_text>
##   ..@ family      : NULL
##   ..@ face        : NULL
##   ..@ italic       : chr NA
##   ..@ fontweight  : num NA
##   ..@ fontwidth   : num NA
##   ..@ colour      : NULL
##   ..@ size         : NULL
##   ..@ hjust        : num 1
##   ..@ vjust        : NULL
##   ..@ angle        : num 45
##   ..@ lineheight  : NULL
##   ..@ margin       : NULL
##   ..@ debug        : NULL
##   ..@ inherit.blank: logi FALSE
## @ complete: logi FALSE
## @ validate: logi TRUE
```

El análisis de la matriz educativa revela una marcada tendencia hacia la homogamia educativa en las disoluciones matrimoniales. Se observa una concentración de casos a lo largo de la diagonal principal (donde el nivel educativo del hombre y la mujer coinciden), lo que sugiere que las parejas tienden a formarse —y consecuentemente a divorciarse— dentro de sus mismos estratos académicos. Sin embargo, el hallazgo más crítico desde la perspectiva de la minería de datos es la alta densidad observada en la intersección de los códigos “9” (esquina superior derecha, color rojo intenso). Dado que este código corresponde a valores “Ignorados” o “No indicados” en los estándares del INE, esto evidencia un sesgo de información faltante significativo; es decir, en una gran proporción de los divorcios registrados, no se capturó el nivel educativo de ninguno de los cónyuges, lo cual limita la caracterización sociodemográfica precisa de este subgrupo.

Relacion: Distribución de la Diferencia de Edad en Divorcios

Diferencia de Edad en Parejas que se Divorcian

Valores positivos: Hombre mayor. Valores negativos: Mujer mayor.



El histograma de distribución de edades en los divorcios muestra una clara asimetría positiva (sesgo hacia la derecha). La mayor concentración de datos se agrupa alrededor de la línea cero, indicando que una gran parte de las disoluciones matrimoniales ocurre en parejas con edades similares (homogamia etaria). Sin embargo, la cola derecha de la distribución es notablemente más larga y densa que la izquierda, lo que revela un patrón social predominante: en las parejas con diferencias de edad significativas que llegan al divorcio, es mucho más frecuente que el hombre sea mayor que la mujer (barras púrpuras extendiéndose hacia los valores positivos). Por el contrario, los casos donde la mujer es considerablemente mayor que el hombre (valores negativos) son estadísticamente menos frecuentes.

Análisis Exploratorio

Estadística Descriptiva Detallada

```
## # A tibble: 2 x 9
##   Variable Media Mediana Desv_Estandar Varianza Minimo Maximo Rango_Intercuartil
##   <chr>     <dbl>    <dbl>        <dbl>    <dbl>    <dbl>    <dbl>           <dbl>
## 1 vic_edad~ 34.2      31       15.0     224.      1      99          16
## 2 differen~  3.16      2        6.06     36.8     -41      75            6
## # i 1 more variable: N_Validos <int>
```

Procedimiento

Para el análisis descriptivo, se programó una función en R que extrajo medidas de tendencia central (media y mediana), dispersión (desviación estándar y rango intercuartílico) y valores extremos. Lo importante fue la depuración de los datos, donde convertimos los códigos de error del INE (como el 999 en edad) en valores NA. Esto evitó que los promedios se dispararan falsamente y permitió que el análisis se basara únicamente en registros válidos, garantizando que la desviación estándar refleje la variabilidad real del fenómeno y no el ruido de la base de datos original.

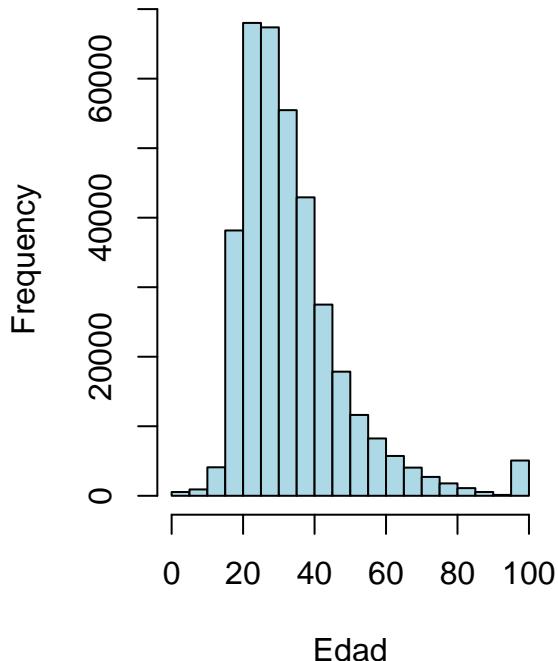
Hallazgos

Los resultados revelan una asimetría positiva en la edad de las víctimas, con una mediana de 31 años; esto significa que el grueso de los casos ocurre en adultos jóvenes, aunque existen víctimas de la tercera edad que elevan la media a 34 años. En cuanto a los divorcios, se detectó una diferencia de edad promedio de 3.16 años (con el hombre siendo mayor), pero lo más impactante fue hallar outliers extremos (diferencias de hasta 75 años). Estos hallazgos confirman que, aunque hay un patrón de comportamiento estándar, existen nichos de vulnerabilidad y dinámicas de pareja muy asimétricas que requieren una atención especial en los modelos de agrupamiento.

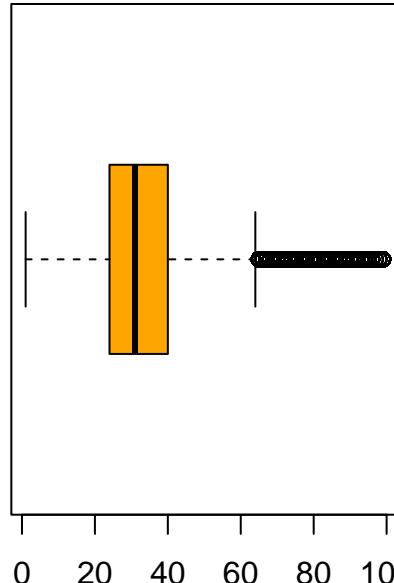
Gráficos Exploratorios

Análisis del dataset de Violencia

Histograma Edad Víctima



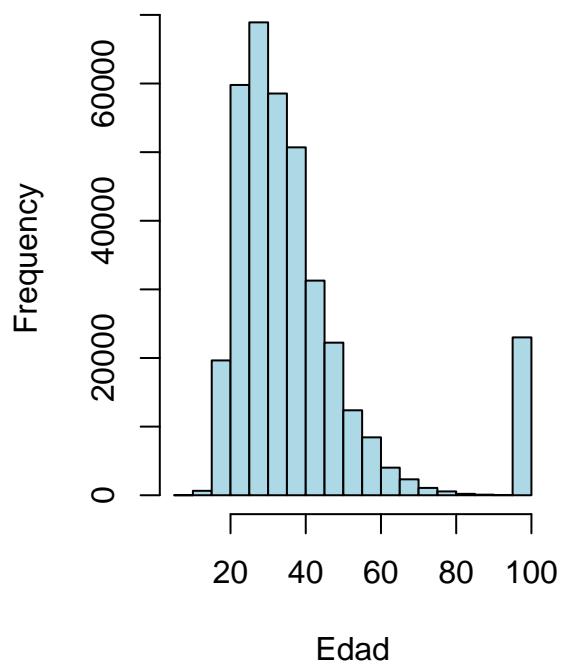
Boxplot Edad Víctima



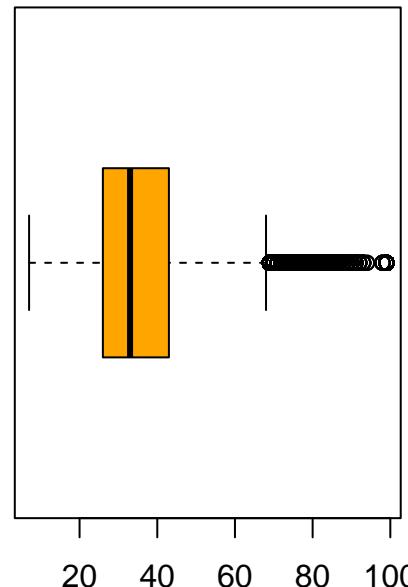
Edad de la Víctima

Al ver las gráficas, lo primero que notamos es que la violencia se concentra muchísimo en los adultos jóvenes, especialmente entre los 20 y 35 años. La ‘cola’ larga hacia la derecha nos indica que, aunque es menos común, también hay víctimas de la tercera edad (esos puntos negros o outliers que aparecen después de los 65 años). Un detalle clave que encontramos es ese pico extraño al final, justo en los 99 años. Eso no significa que haya muchas víctimas de esa edad, sino que es un código que usan para decir ‘edad desconocida’. Es importante haberlo detectado visualmente, porque si no lo limpiamos, nos va a inflar el promedio artificialmente.

Histograma Edad Agresor

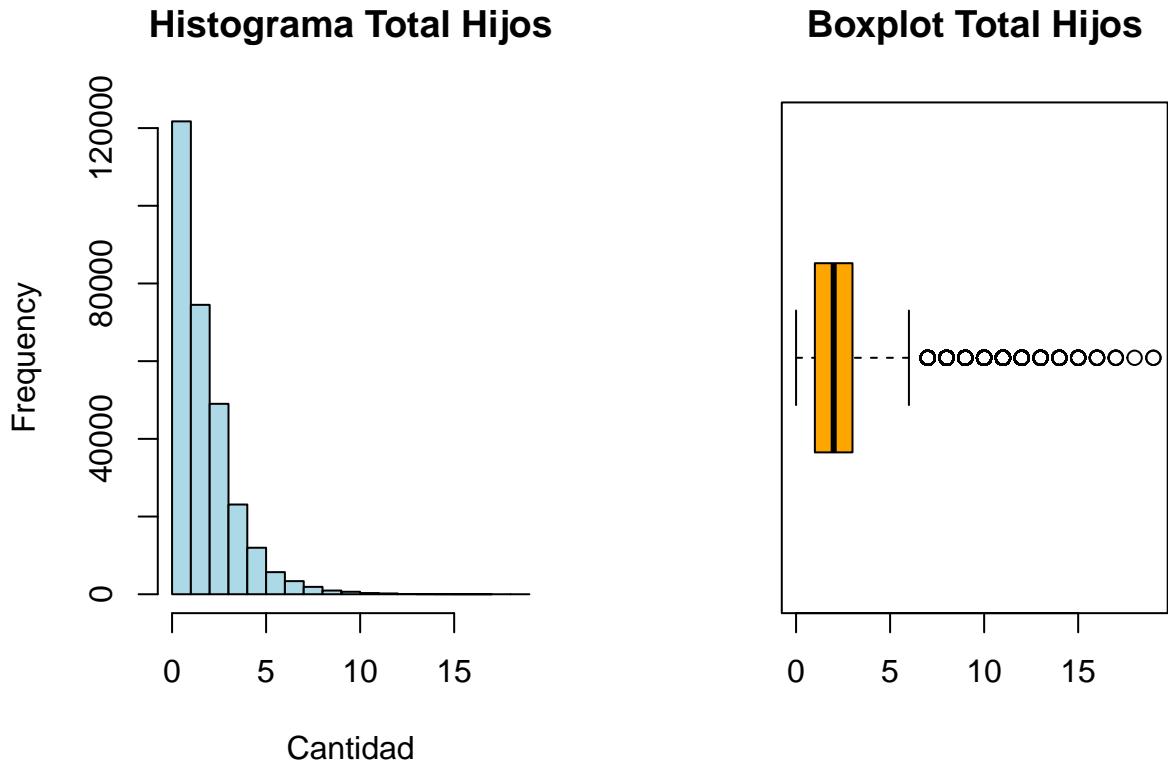


Boxplot Edad Agresor



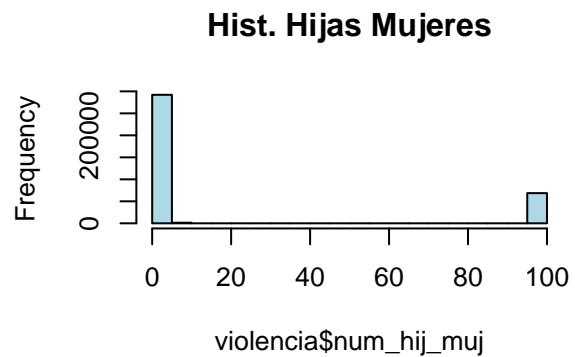
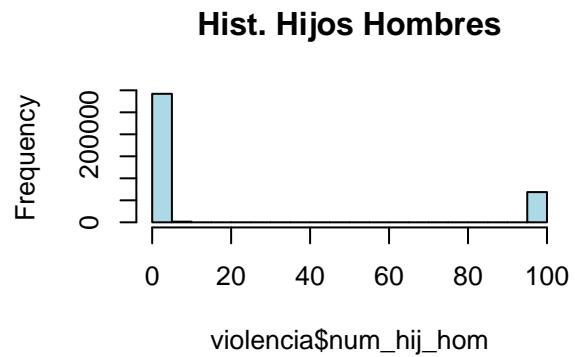
Edad del Agresor

Lo primero que salta a la vista es que la gran mayoría de los casos se concentran en familias pequeñas, de entre 0 y 3 hijos. El histograma tiene una caída muy brusca, lo que nos confirma que tener muchos hijos es algo poco común en este grupo. Por otro lado, el diagrama de caja nos muestra una fila larga de puntos negros a la derecha. Estos son los valores atípicos (outliers). Ojo, no necesariamente son errores de datos, estos representan familias reales que tienen 8, 10 o hasta 15 hijos, pero que estadísticamente se comportan muy diferente al promedio general.



Total de Hijos

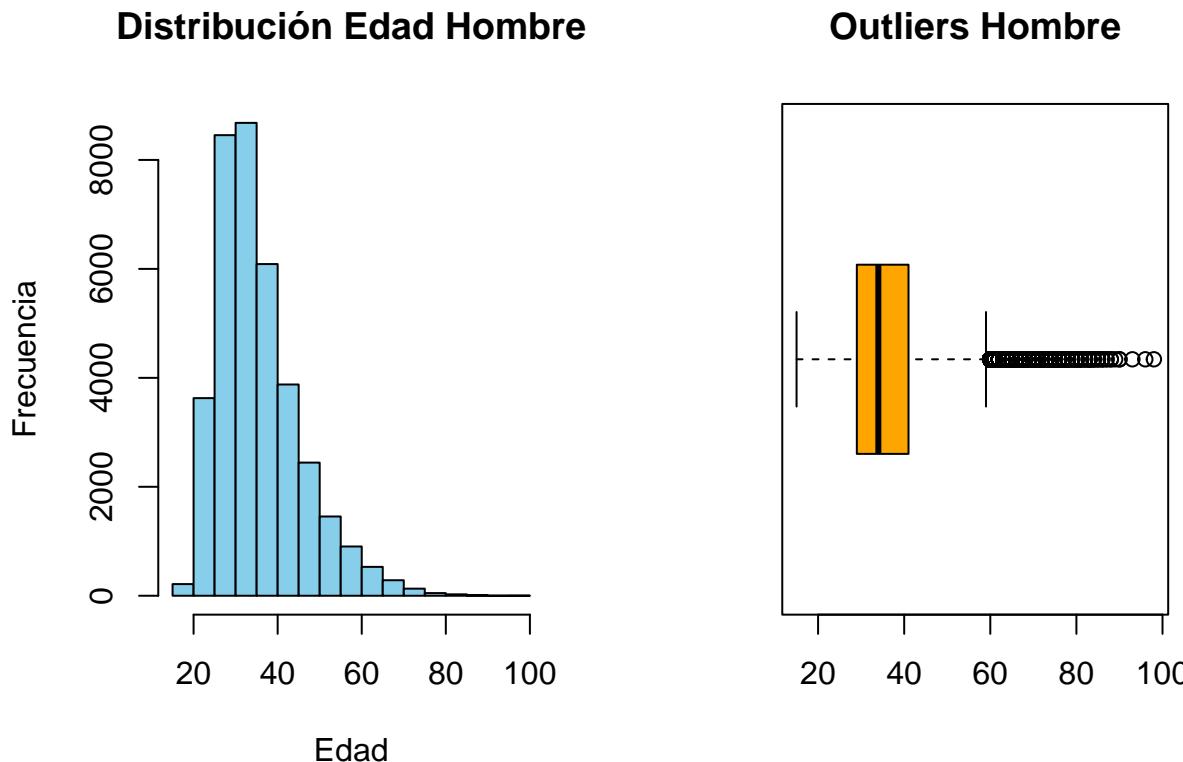
Al ver estas gráficas, es evidente que la gran mayoría de los casos involucran familias muy pequeñas, de entre 0 y 3 hijos. El histograma cae en picada, lo que nos confirma que tener una familia numerosa es la excepción y no la regla en este grupo. Por otro lado, el diagrama de caja nos muestra una larga fila de puntos negros hacia la derecha. Estos son los valores atípicos, pero ojo, no son errores, sino familias reales numerosas (de 6 hasta más de 15 hijos) que, aunque existen, estadísticamente se salen por completo del comportamiento ‘normal’ o promedio de la población estudiada.



Hijos Hombres y Mujeres (Desglosado)

Al separar los datos entre hijos hombres y mujeres, confirmamos un patrón técnico muy importante. Si observan los histogramas, la inmensa mayoría se agrupa en el cero o números bajos, pero vuelve a aparecer esa barra extraña al final, cerca del 100. Los diagramas de caja lo hacen evidente: ese punto solitario allá arriba no representa a una familia con 99 hijos varones, sino que es indudablemente un código de dato no registrado. Este es un hallazgo clave de limpieza, porque si no filtramos esos '99', cualquier cálculo de promedio de hijos saldrá totalmente erróneo.

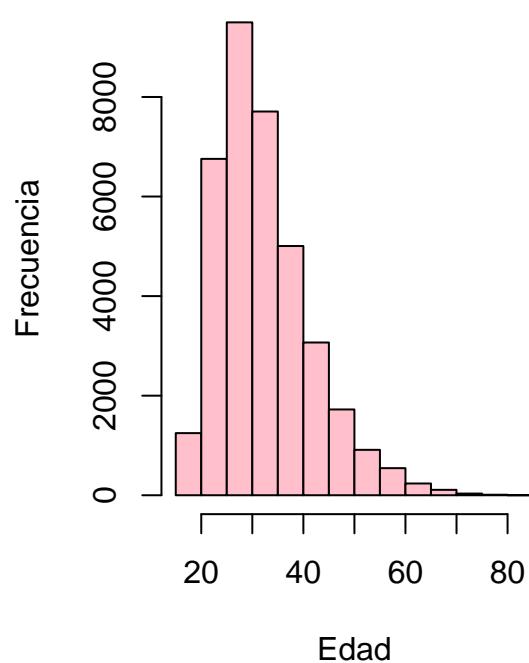
Análisis del dataset de Divorcios



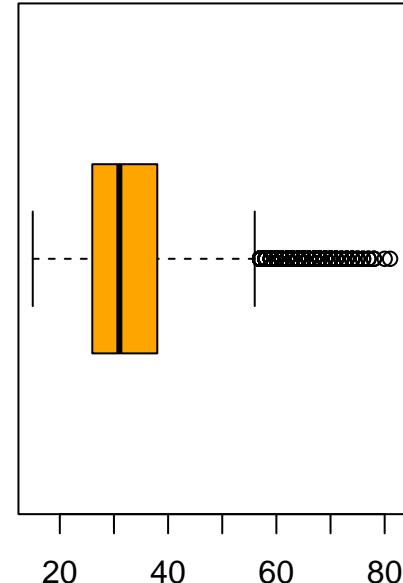
Edad del Hombre

Al analizar la edad de los hombres al momento del divorcio, vemos claramente que es un fenómeno concentrado en la adultez media. El histograma tiene su pico más alto entre los 30 y 40 años, lo que nos indica que esa es la etapa más crítica para las disoluciones matrimoniales en este grupo. Por otro lado, el diagrama de caja nos cuenta la historia completa de que aunque el grueso de los casos está en esa franja joven, la fila de puntos negros a la derecha (los outliers) nos confirma que el divorcio no tiene edad límite. Existen casos registrados de hombres de 70, 80 y hasta casi 100 años separándose, lo cual es estadísticamente atípico pero socialmente real.

Distribución Edad Mujer



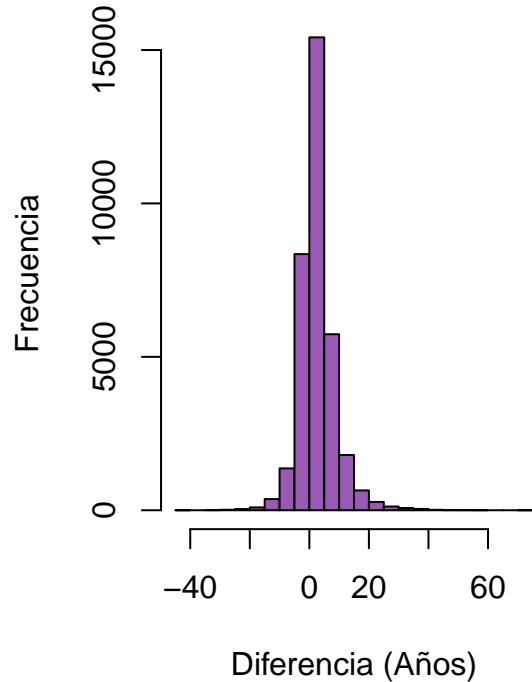
Outliers Mujer



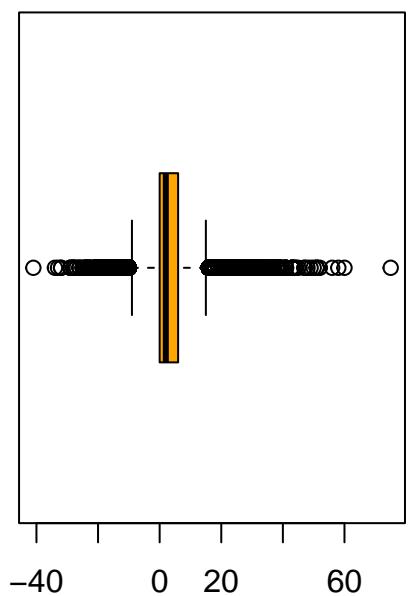
Edad de la Mujer

Al ver las gráficas de las mujeres, notamos un patrón muy similar al de los hombres, pero con una diferencia clave: el pico de divorcios ocurre un poco antes. El histograma nos grita que la mayor concentración de casos está entre los 25 y 35 años, confirmando que las mujeres tienden a pasar por este proceso siendo ligeramente más jóvenes que sus exparejas. El diagrama de caja confirma esta tendencia con una mediana baja, pero también nos muestra una larga cola de outliers que llega hasta los 80 años. Esto significa que, aunque la estadística dice que el divorcio es cosa de jóvenes, hay un grupo real y visible de mujeres de la tercera edad que también están disolviendo sus matrimonios.

Brecha de Edad (Hombre – Mujer)



Detectando Brechas Extremas



Diferencia de edad

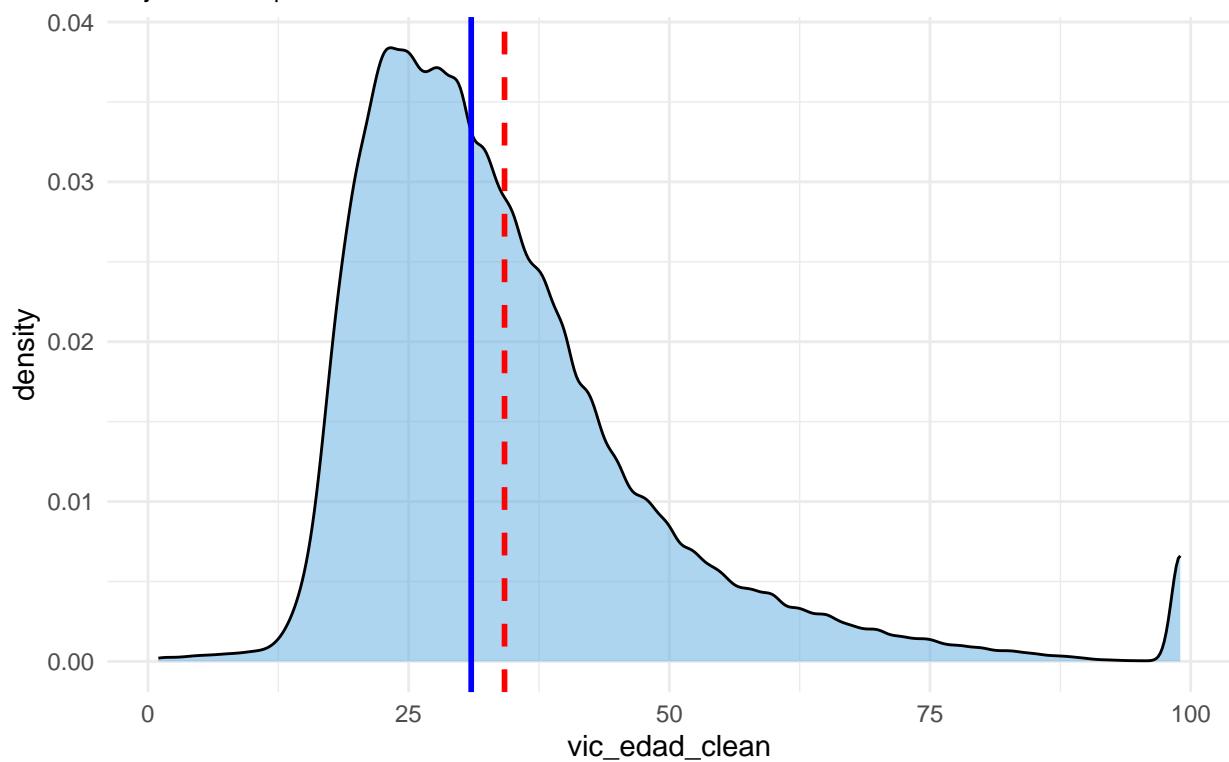
Lo primero que salta a la vista es que la gran mayoría de las parejas tienen edades muy similares. Ese pico gigante y puntiagudo en el centro del histograma nos confirma que lo estándar es divorciarse de alguien de tu misma generación, con una diferencia de apenas 0 a 5 años. Pero lo más interesante está en el diagrama de caja, por ejemplo fíjense en la cantidad brutal de puntos negros a los lados. Esos son los outliers extremos. Tenemos casos donde el hombre es hasta 70 años mayor que su pareja (lado derecho), y otros donde la mujer es 40 años mayor que él (lado izquierdo). Aunque son casos raros, existen y rompen por completo la regla de la edad similar.

Distribución de variables numéricos

Análisis del dataset de violencia

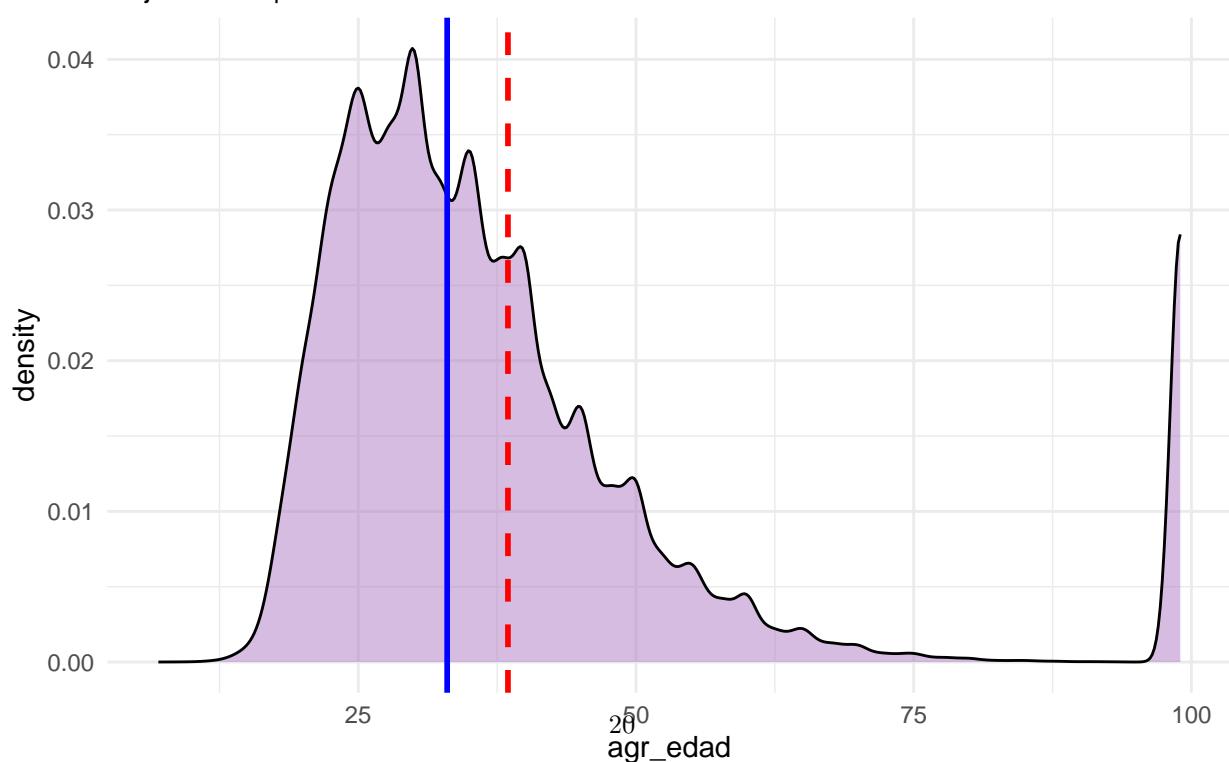
Distribución: Edad Víctima

Roja=Media | Azul=Mediana



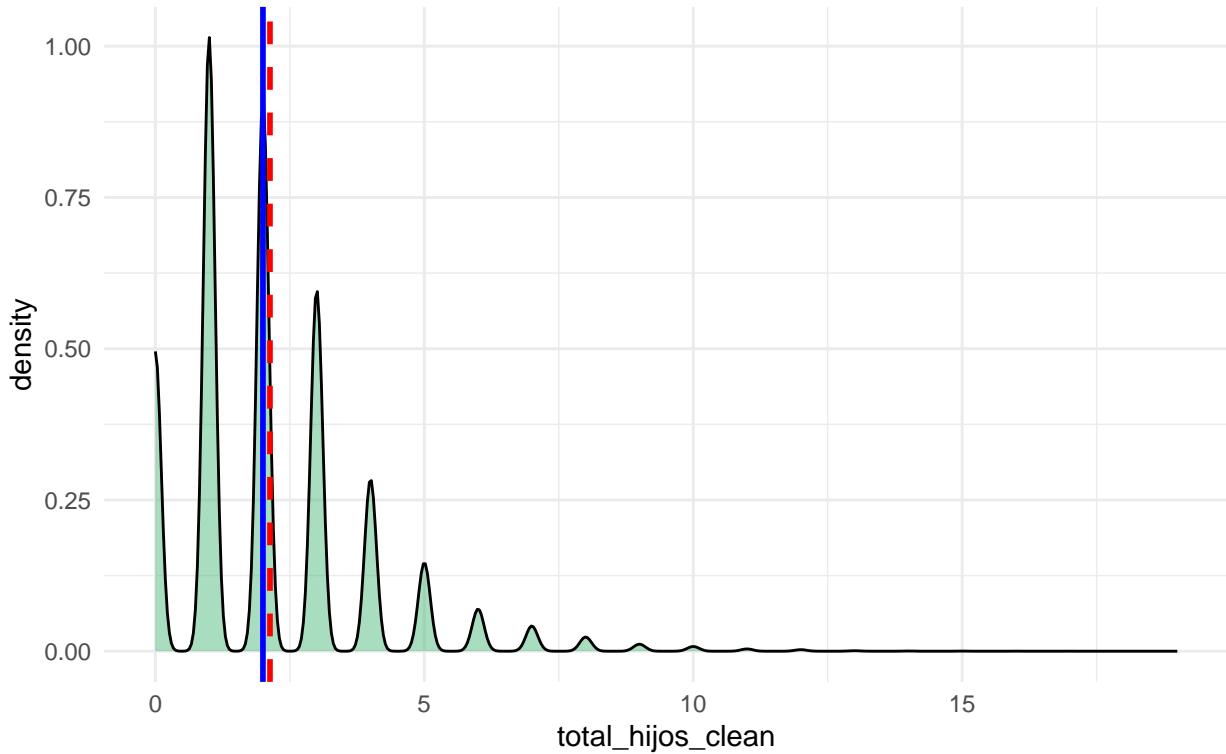
Distribución: Edad Agresor

Roja=Media | Azul=Mediana



Distribución: Total Hijos

Roja=Media | Azul=Mediana



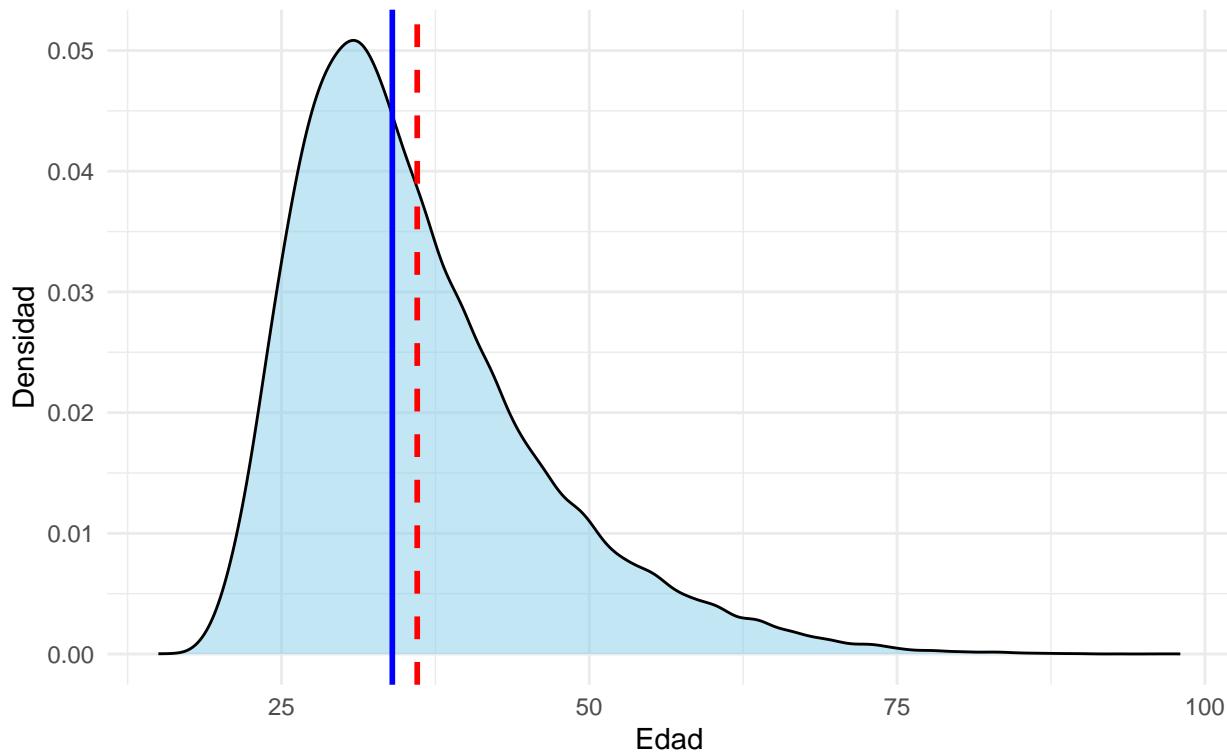
Edad de la víctima: Al observar este gráfico, la prueba de que no existe normalidad es clara ya que las líneas no coinciden, por ejemplo la línea azul (Mediana) nos marca el punto real donde se divide la población (alrededor de los 30 años), mientras que la línea roja (Media) está desplazada hacia la derecha, pero ¿por qué pasa esto? Por la ‘cola’ de datos que vemos a la derecha y, sobre todo, por ese pequeño pico extraño al final (cerca de los 100 años). Esos valores extremos están ‘inflando’ el promedio artificialmente. Por eso, para este proyecto, la Mediana será un dato mucho más confiable que el promedio para describir a la víctima típica.

Edad del agresor: Aquí el fenómeno es aún más evidente porque vemos nuevamente que la línea roja (Media) se separa de la azul hacia la derecha, confirmando un sesgo positivo. Pero lo más importante para la limpieza de datos es notar cómo la curva sube repentinamente al final del gráfico (a la derecha del todo). Eselevantamiento es causado por los códigos ‘99’ (edad desconocida). Al estar esos datos ahí, están arrastrando la línea roja, haciendo parecer que la edad promedio de los agresores es mayor de lo que realmente es por esto debemos limpiar esos ‘99’ antes de modelar.

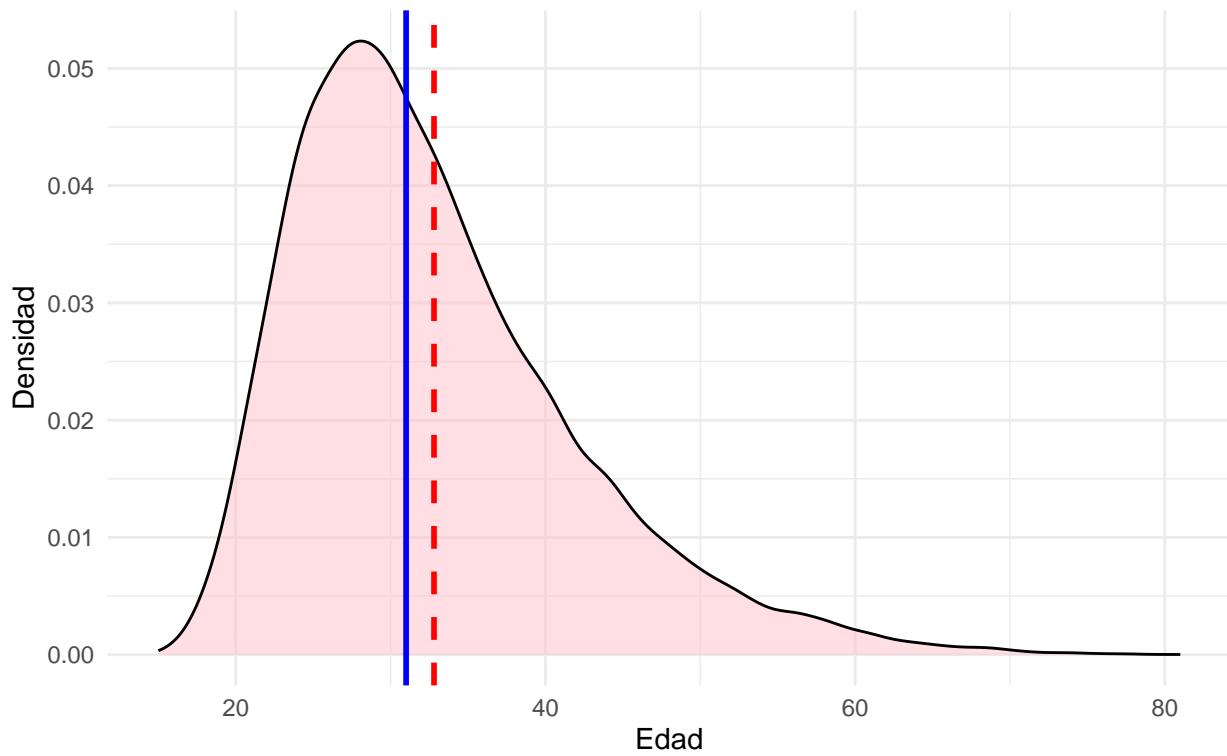
Total hijos: Esta gráfica tiene forma de sierra pero lo crítico es la posición de las líneas. La línea azul (Mediana) está clavada en 2 hijos, que es la realidad más común de las familias en la base de datos. Sin embargo, la línea roja (Media) está corrida a la derecha. Esto sucede porque las familias ‘atípicas’ con 8, 10 o 15 hijos pesan mucho en el cálculo matemático, moviendo el promedio, al final podemos ver que la variable no es normal y está fuertemente sesgada ya que al usar el promedio aquí nos daría un número decimal que no representa la realidad del hogar típico.

Análisis del dataset de divorcios

Distribución: Edad Hombre
Línea Roja (Media) vs Azul (Mediana)

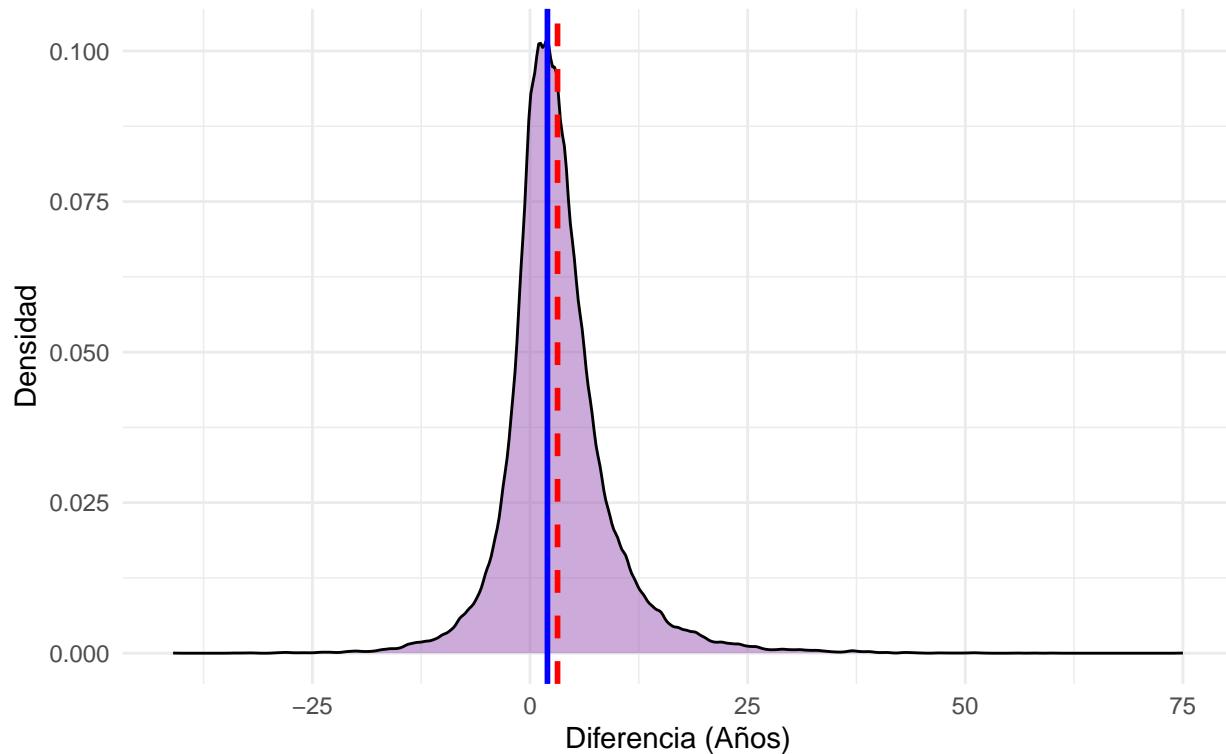


Distribución: Edad Mujer
Línea Roja (Media) vs Azul (Mediana)



Distribución: Diferencia de Edad

Línea Roja (Media) vs Azul (Mediana)



Edad del hombre: Al analizar esta gráfica, la falta de normalidad es evidente. La línea azul (Mediana) nos muestra que el hombre ‘típico’ se divorcia cerca de los 38 años. Sin embargo, la línea roja (Media) aparece desplazada hacia la derecha, esto nos dice que tenemos una asimetría positiva. Aunque el grueso de los divorcios ocurre en la adultez media, hay suficientes hombres de 60, 70 o más años divorciándose como para inflar el promedio matemático, por eso, si solo viéramos el promedio, creeríamos que la gente se divorcia más vieja de lo que realmente es lo común.

Edad de la mujer: Aquí confirmamos que el comportamiento es casi idéntico al de los hombres, pero ocurriendo un poco antes. La línea azul se planta firmemente alrededor de los 35 años, marcando la realidad de la mayoría, después nuevamente, la línea roja (Media) huye hacia la derecha. Esto ratifica estadísticamente que los casos de mujeres mayores distorsionan la media. Al igual que con los hombres, la distribución no es normal, y usar métodos estadísticos que asuman una Campana de Gauss perfecta podría llevarnos a errores.

Diferencia de edad: Esta es la gráfica más trampa de todas porque a primera vista parece una campana perfecta porque las líneas roja y azul están casi pegadas. Pero si miramos bien, la forma es demasiado puntiaguda (lo que llamamos leptocúrtica). Esto significa que hay una concentración excesiva de datos en el centro porque la inmensa mayoría de las parejas se lleva menos de 5 años, por otro lado las ‘colas’ largas a los lados (las partes bajas y planas de la curva) nos indican que existen valores extremos—parejas con 20 o 30 años de diferencia que rompen la estructura de una distribución normal estándar, aunque el promedio sea acertado aquí, la variabilidad es engañosa.

```
## # A tibble: 59 x 5
##   var           dataset     n   n_na  pct_na
##   <chr>         <chr>    <int> <int>   <dbl>
## 1 articulotras1 violencia 363888 362306 99.6
## 2 articulotras2 violencia 363888 362306 99.6
## 3 articulotras3 violencia 363888 362306 99.6
## 4 articulotras4 violencia 363888 362306 99.6
```

```

## 5 articulocodpen1      violencia 363888 361613 99.4
## 6 articulocodpen2      violencia 363888 361613 99.4
## 7 articulocodpen3      violencia 363888 361613 99.4
## 8 articulocodpen4      violencia 363888 361613 99.4
## 9 tipo_discaq         violencia 363888 356195 97.9
## 10 filter               violencia 363888 333900 91.8
## 11 agr_grupet          violencia 363888 327781 90.1
## 12 articulovcm1         violencia 363888 324944 89.3
## 13 articulovcm2         violencia 363888 324944 89.3
## 14 articulovcm3         violencia 363888 324944 89.3
## 15 articulovcm4         violencia 363888 324944 89.3
## 16 inst_donde_denuncio violencia 363888 318892 87.6
## 17 numero_boleta        violencia 363888 290259 79.8
## 18 agr_dedica           violencia 363888 288194 79.2
## 19 articulovif1          violencia 363888 261940 72.0
## 20 articulovif2          violencia 363888 261940 72.0
## 21 articulovif3          violencia 363888 261940 72.0
## 22 articulovif4          violencia 363888 261940 72.0
## 23 organismo_jurisdiccional violencia 363888 260643 71.6
## 24 conductor            violencia 363888 260643 71.6
## 25 organismo_remiter     violencia 363888 260643 71.6
## 26 vic_ocup              violencia 363888 242377 66.6
## 27 tipo_medida           violencia 363888 206085 56.6
## 28 medidas_seguridad     violencia 363888 161992 44.5
## 29 ley_aplicable          violencia 363888 160694 44.2
## 30 vic_dedica            violencia 363888 122770 33.7
## 31 agr_ocup              violencia 363888 76353 21.0
## 32 total_hijos_clean     violencia 363888 70366 19.3
## 33 agr_gurpet             violencia 363888 36107 9.92
## 34 agresores_otros_total violencia 363888 36107 9.92
## 35 agr_otros_hom          violencia 363888 36107 9.92
## 36 agr_otras_muj          violencia 363888 36107 9.92
## 37 agr_otros_nos          violencia 363888 36107 9.92
## 38 agr_otras_nas          violencia 363888 36107 9.92
## 39 mes_nombre              violencia 363888 4013 1.10
## 40 total_hijos            violencia 363888 1694 0.466
## 41 num_hij_hom             violencia 363888 1694 0.466
## 42 num_hij_muj             violencia 363888 1694 0.466
## 43 vic_est_civ            violencia 363888 1694 0.466
## 44 vic_alfab              violencia 363888 689 0.189
## 45 vic_escolaridad         violencia 363888 689 0.189
## 46 vic_trabaja            violencia 363888 689 0.189
## 47 agr_est_civ             violencia 363888 26 0.00715
## 48 GETHOM                  divorcios 71576 66419 92.8
## 49 GETMUJ                  divorcios 71576 66419 92.8
## 50 OCUHOM                  divorcios 71576 66419 92.8
## # i 9 more rows

## # A tibble: 1 x 6
##   out_vic_iqr out_agr_iqr out_hij_iqr out_vic_mad out_agr_mad out_hij_mad
##   <int>       <int>       <int>       <int>       <int>       <int>
## 1    16606     25852      7503      23100      30680      13194

## # A tibble: 15 x 8

```

```

##      cod_depto hec_mes vic_edad_clean agr_edad_clean total_hijos_clean out_vic_iqr
##      <dbl>     <dbl>          <dbl>          <dbl>          <dbl> <lgl>
## 1       1         1           99            37            2 TRUE
## 2       1         1           99            28            2 TRUE
## 3       1         1           99            25           NA TRUE
## 4       1         1           99            27           NA TRUE
## 5       1        99           99            24           NA TRUE
## 6      99        99           99            39            3 TRUE
## 7       1         1           99            33           NA TRUE
## 8       1         1           99            22            1 TRUE
## 9       1         1           99            22           NA TRUE
## 10      1         1           99            99           NA TRUE
## 11      3         1           99            51           NA TRUE
## 12      7         1           99            25           NA TRUE
## 13      7         1           99            22           NA TRUE
## 14      9         1           99            22           NA TRUE
## 15      9         1           99            99           NA TRUE
## # i 2 more variables: out_agr_iqr <lgl>, out_hij_iqr <lgl>

## # A tibble: 15 x 8
##      cod_depto hec_mes vic_edad_clean agr_edad_clean total_hijos_clean out_vic_iqr
##      <dbl>     <dbl>          <dbl>          <dbl>          <dbl> <lgl>
## 1       7         2           1           99           NA FALSE
## 2      12        12          1           22           NA FALSE
## 3      22        12          1           21           NA FALSE
## 4       7        12          1           22           NA FALSE
## 5      16        10          1           20           NA FALSE
## 6       9        11          1           20           NA FALSE
## 7      18        11          1           18           NA FALSE
## 8      11        3           1           22           NA FALSE
## 9      16        5           1           15           NA FALSE
## 10      2         7           1           25           NA FALSE
## 11      16        8           1           20           NA FALSE
## 12      16        11          1           23           NA FALSE
## 13      17        2           1           23           NA FALSE
## 14      15        1           1           33           NA FALSE
## 15       5        4           1           20           NA FALSE
## # i 2 more variables: out_agr_iqr <lgl>, out_hij_iqr <lgl>

##      vic_edad_clean agr_edad_clean total_hijos_clean
## vic_edad_clean      1.0000000   0.2802643   0.3681686
## agr_edad_clean      0.2802643   1.0000000   0.1163771
## total_hijos_clean   0.3681686   0.1163771   1.0000000

##      vic_edad_clean agr_edad_clean total_hijos_clean
## vic_edad_clean      1.0000000   0.4663399   0.4291053
## agr_edad_clean      0.4663399   1.0000000   0.2342243
## total_hijos_clean   0.4291053   0.2342243   1.0000000

##      edad_hom_clean edad_muj_clean diferencia_edad
## edad_hom_clean      1.0000000   0.7923713   0.4525175
## edad_muj_clean      0.7923713   1.0000000  -0.1854439
## diferencia_edad     0.4525175  -0.1854439   1.0000000

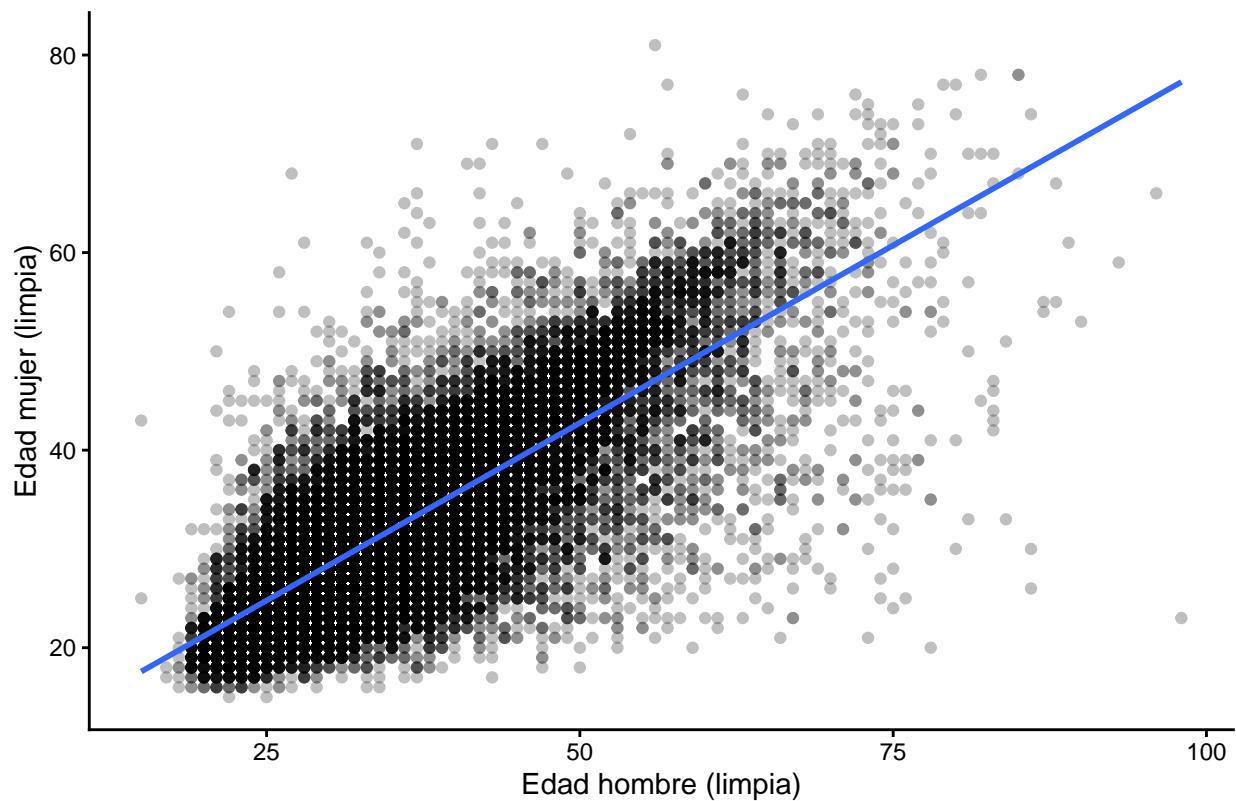
```

```

##          edad_hom_clean edad_muj_clean diferencia_edad
## edad_hom_clean      1.0000000    0.7785327   0.3661467
## edad_muj_clean      0.7785327    1.0000000  -0.2154773
## diferencia_edad     0.3661467   -0.2154773   1.0000000

```

Edad hombre vs mujer en divorcios



```

##          violencia_total divorcios_total
## violencia_total      1.000000    0.433211
## divorcios_total      0.433211    1.000000

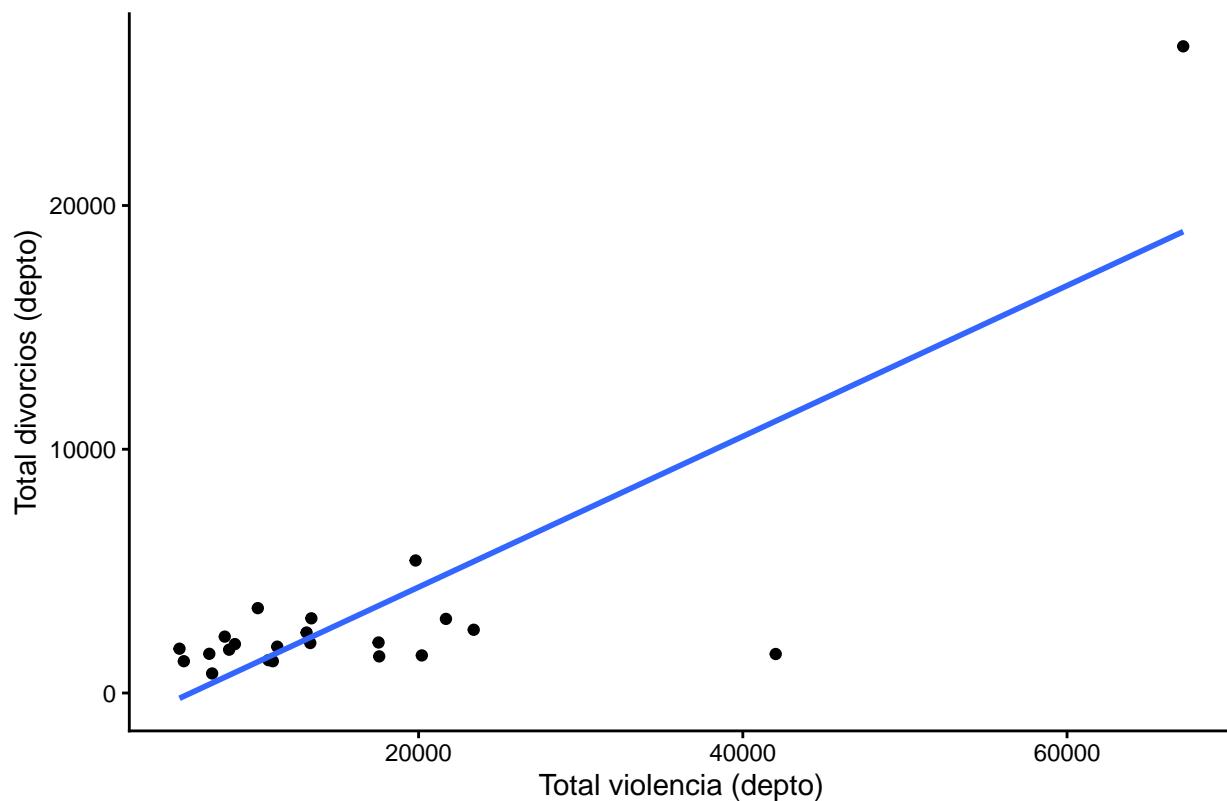
```

```

##          violencia_total divorcios_total
## violencia_total      1.000000    0.8171023
## divorcios_total      0.8171023   1.0000000

```

Violencia intrafamiliar vs divorcios por departamento



Se identificó que el valor 99 corresponde a la categoría 'Ignorado', por lo que constituye un valor faltante codificado. Su presencia generaba acumulaciones artificiales en los diagramas de dispersión (bandas cercanas a 99) y podía distorsionar medidas de asociación, especialmente Pearson. Por ello, se decidió recodificar 99 como NA (y excluir cod_dept = 99 del análisis por departamento), lo cual permite interpretar correctamente las correlaciones: en violencia intrafamiliar se observa una asociación positiva entre la edad de la víctima y del agresor, y una relación moderada entre edad de la víctima y número de hijos; en divorcios, las edades del hombre y la mujer se correlacionan fuertemente; y a nivel departamental, violencia y divorcios presentan una relación monotónica alta, aunque la relación lineal es sensible a outliers y diferencias de escala entre departamentos

```
## # A tibble: 2 x 3
##   valor      n    pct
##   <fct>    <int> <dbl>
## 1 Mujer  322092  88.5
## 2 Hombre  41796   11.5

## # A tibble: 4 x 3
##   valor      n    pct
##   <fct>    <int> <dbl>
## 1 Casado(a) 149639  41.1
## 2 Unido(a)  90564  24.9
## 3 Ignorado  73213  20.1
## 4 Soltero(a) 50472  13.9

## # A tibble: 6 x 3
##   valor      n    pct
##   <fct>    <int> <dbl>
```

```

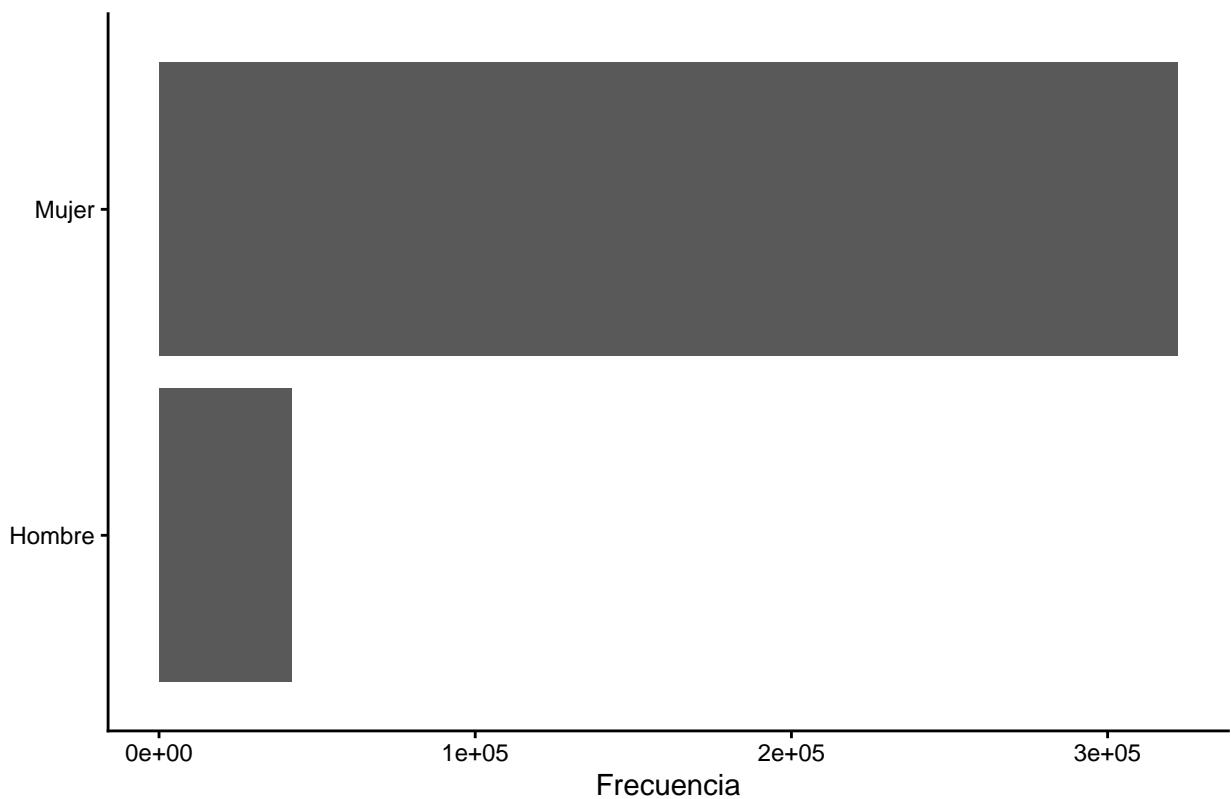
##   <fct>      <int>  <dbl>
## 1 Esposo/Conviviente 110364 30.3
## 2 Ex-Esposo/Ex-Pareja 109926 30.2
## 3 Otro/No especificado 77357 21.3
## 4 Padre/Madre         56559 15.5
## 5 Hijo(a)              8315  2.29
## 6 Hermano(a)          1367  0.376

## # A tibble: 13 x 3
##       valor      n    pct
##       <fct>     <int> <dbl>
## 1 Mayo        32123  8.83
## 2 Julio        31895  8.77
## 3 Agosto       31049  8.53
## 4 Marzo        30965  8.51
## 5 Abril        30609  8.41
## 6 Junio        30323  8.33
## 7 Febrero      29662  8.15
## 8 Enero         29644  8.15
## 9 Septiembre   29385  8.08
## 10 Octubre      29094  8.00
## 11 Noviembre    27967  7.69
## 12 Diciembre    27159  7.46
## 13 Ignorado     4013   1.10

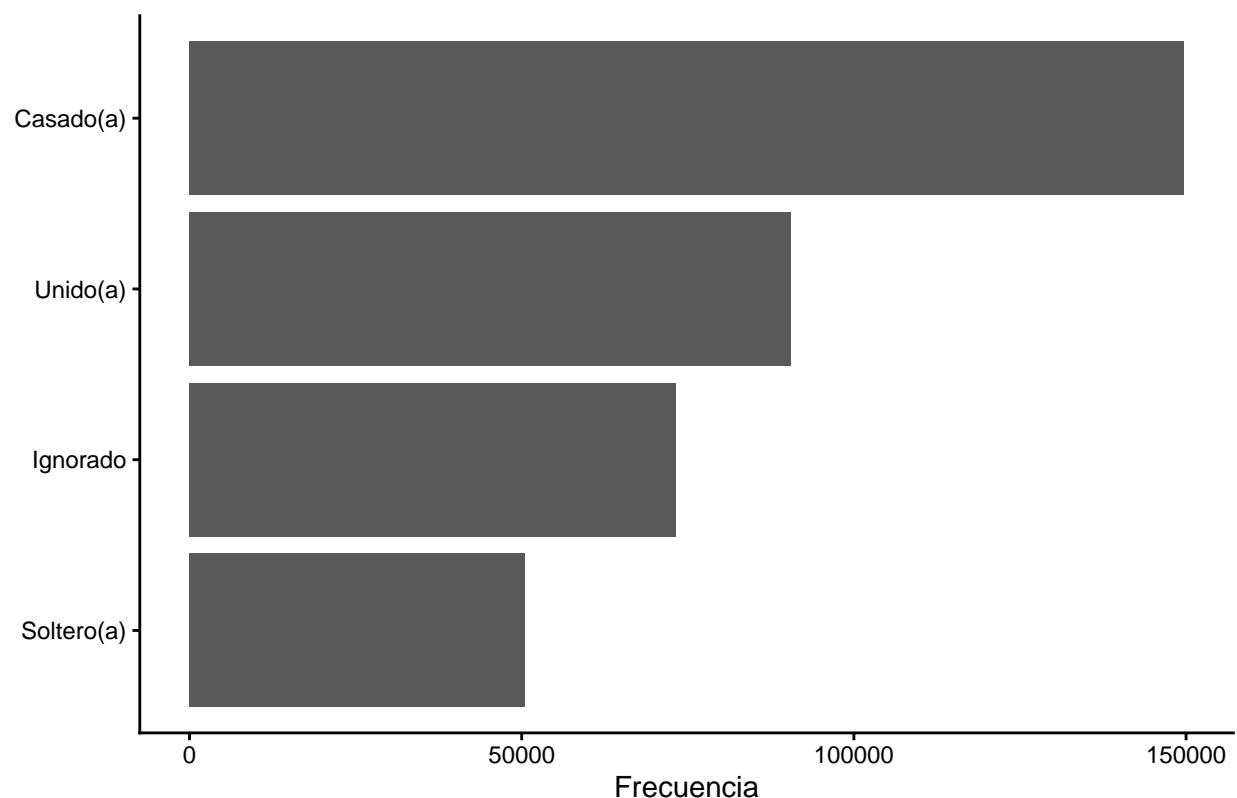
## # A tibble: 23 x 3
##       valor      n    pct
##       <fct>     <int> <dbl>
## 1 1          67166 18.5
## 2 16         42029 11.5
## 3 10         23400  6.43
## 4 12         21688  5.96
## 5 4          20199  5.55
## 6 9          19815  5.45
## 7 3          17571  4.83
## 8 11         17530  4.82
## 9 22         13385  3.68
## 10 6         13318  3.66
## # i 13 more rows

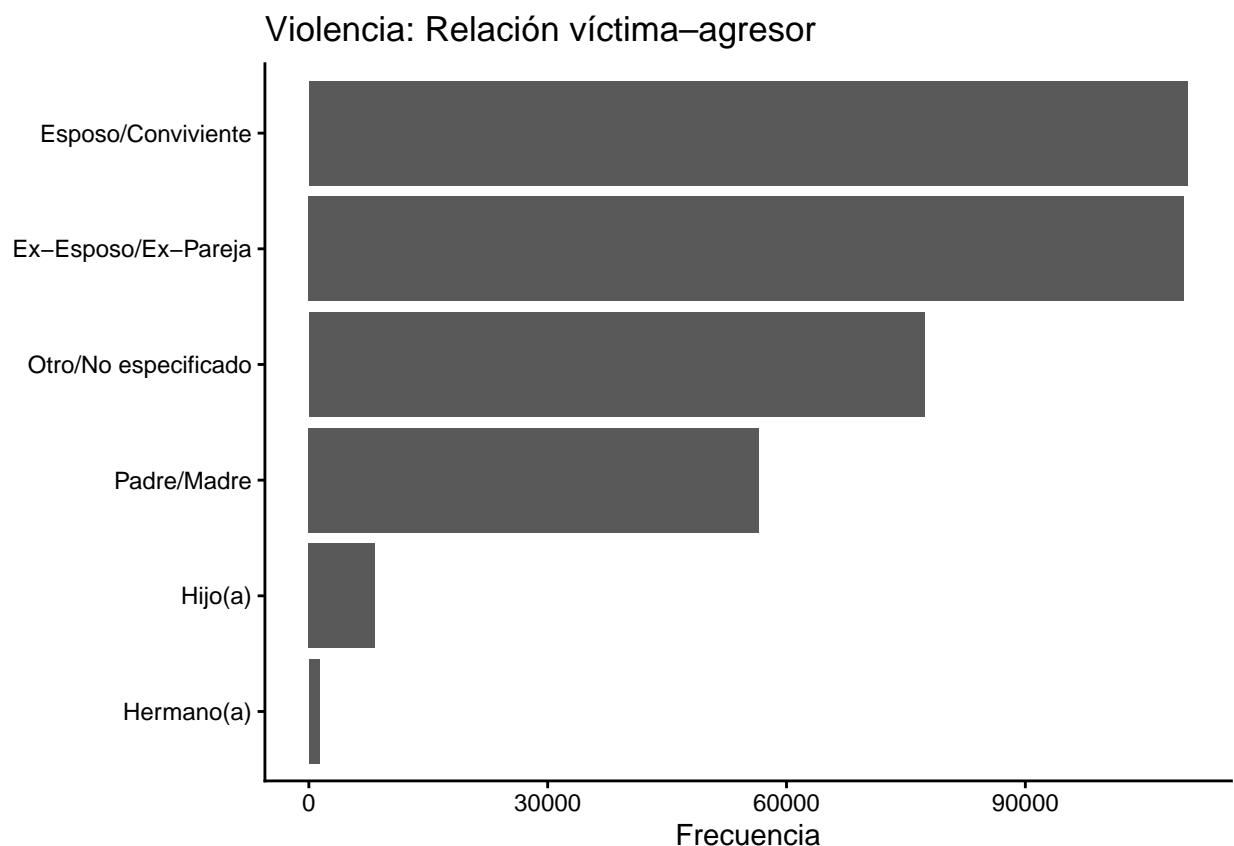
```

Violencia: Sexo de la víctima

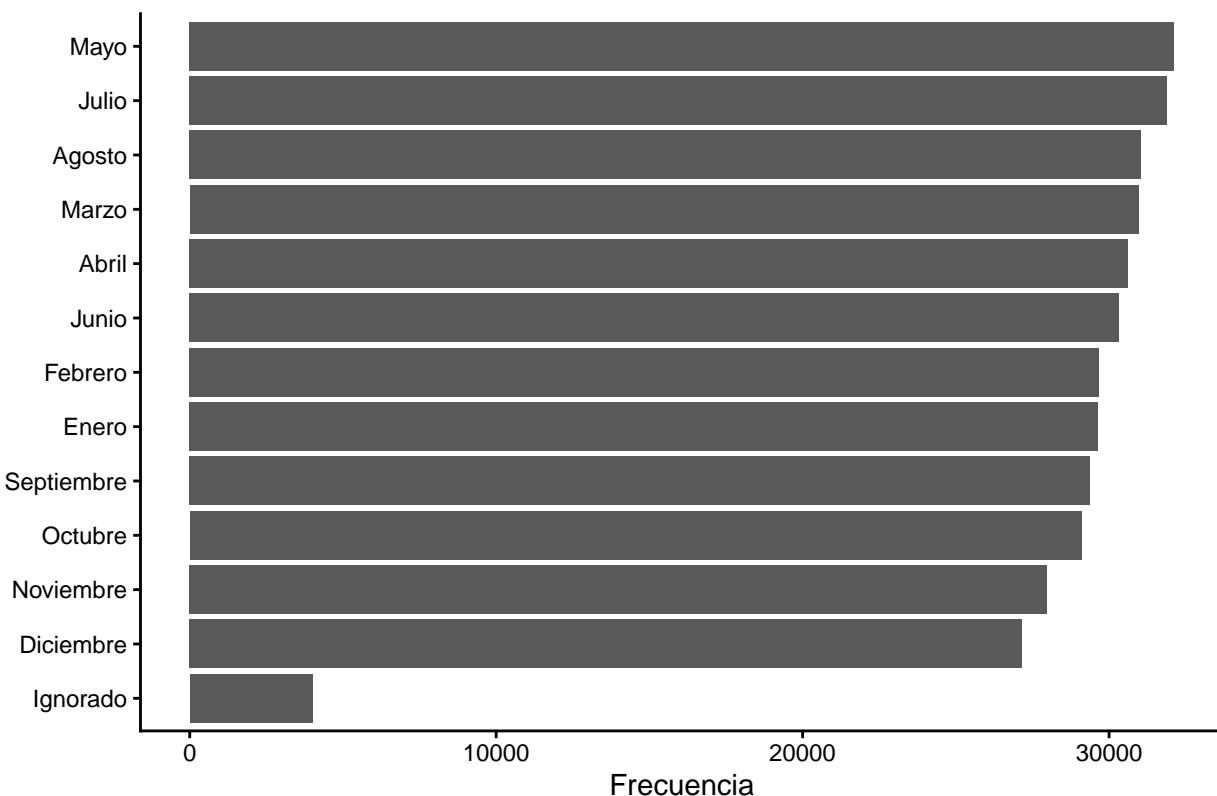


Violencia: Estado civil de la víctima





Violencia: Mes del hecho (incluye Ignorado)

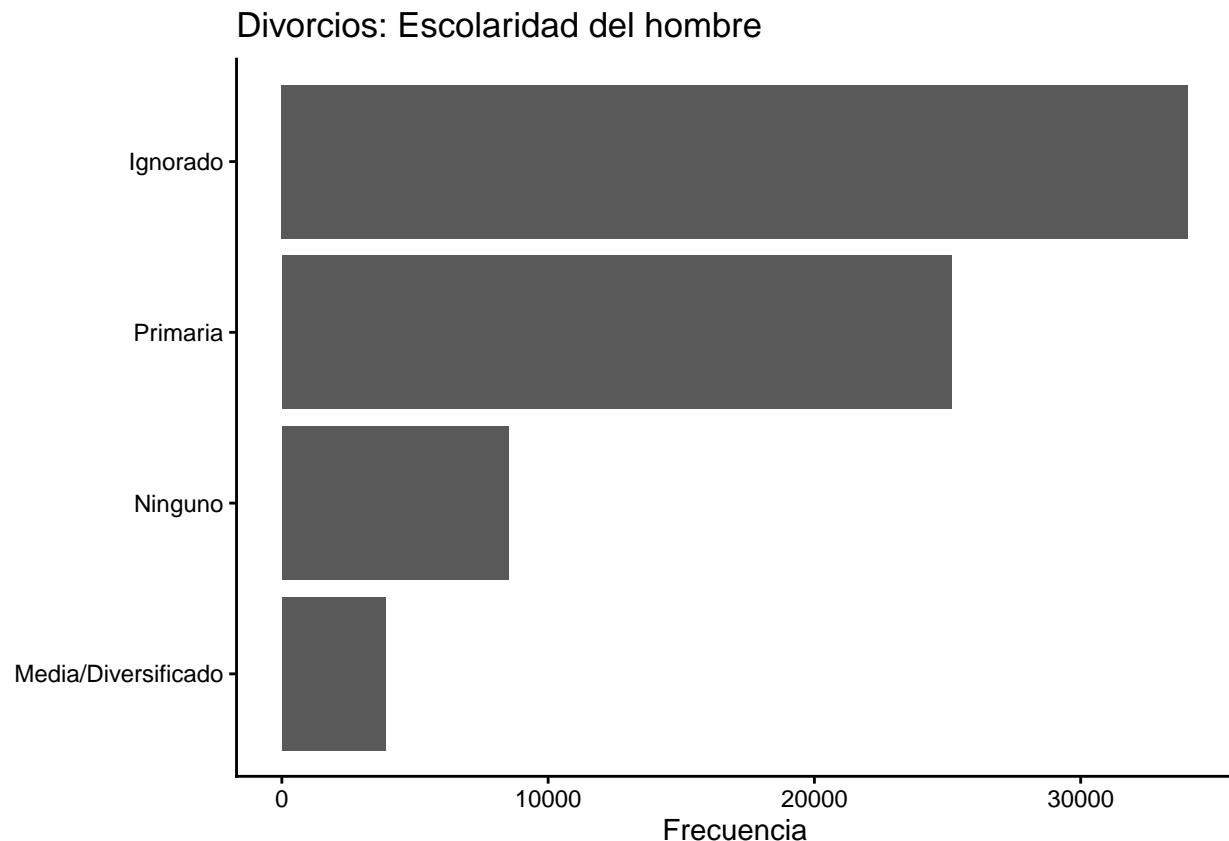


```
## # A tibble: 4 x 3
##   valor           n   pct
##   <fct>      <int> <dbl>
## 1 Ignorado    34033 47.5
## 2 Primaria    25145 35.1
## 3 Ninguno     8517 11.9
## 4 Media/Diversificado 3881  5.42

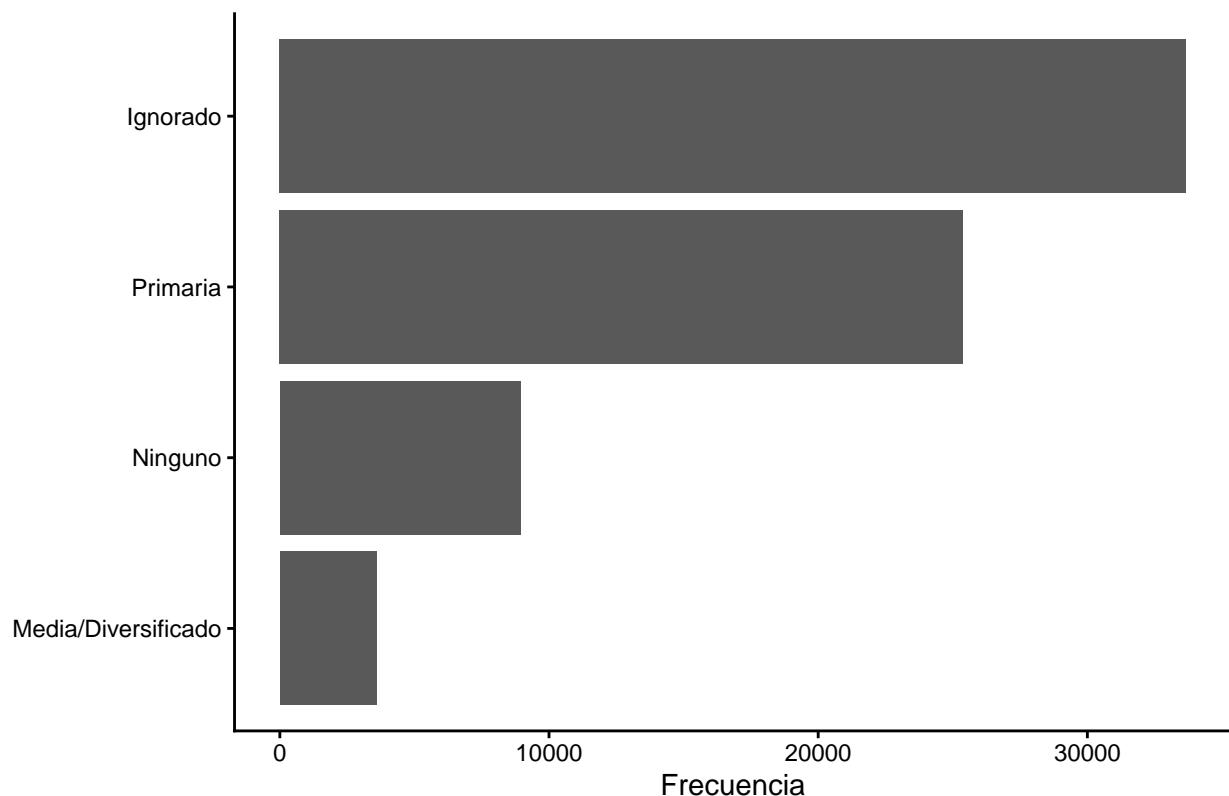
## # A tibble: 4 x 3
##   valor           n   pct
##   <fct>      <int> <dbl>
## 1 Ignorado    33668 47.0
## 2 Primaria    25386 35.5
## 3 Ninguno     8934 12.5
## 4 Media/Diversificado 3588  5.01

## # A tibble: 22 x 3
##   valor      n   pct
##   <fct> <int> <dbl>
## 1 1       26530 37.1
## 2 9       5437  7.60
## 3 5       3481  4.86
## 4 22      3064  4.28
## 5 12      3041  4.25
## 6 10      2594  3.62
```

```
## 7 13      2480  3.46
## 8 18      2313  3.23
## 9 11      2072  2.89
## 10 6      2051  2.87
## # i 12 more rows
```



Divorcios: Escolaridad de la mujer



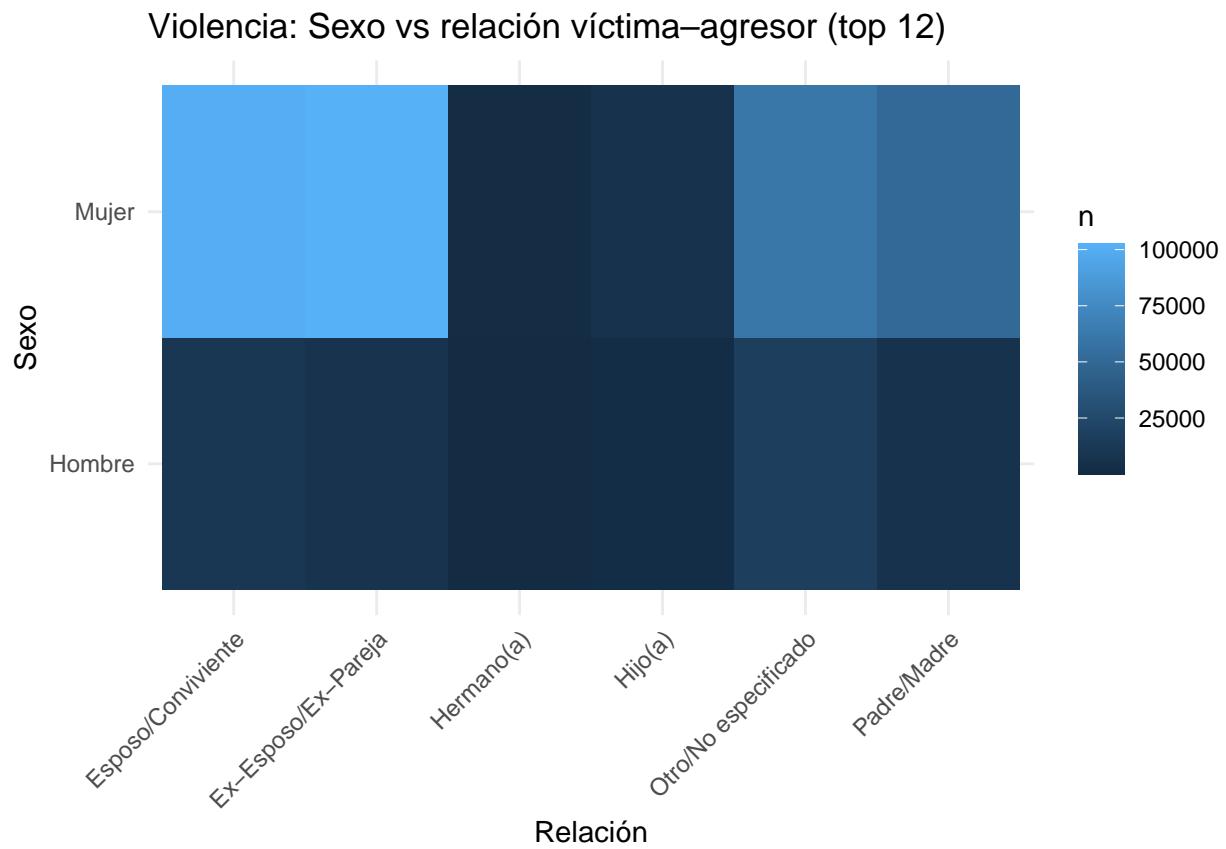
```
## $V
## [1] 0.06274014
##
## $p_value
## [1] 2.770845e-310
##
## $chi2
## [1] 1432.381
##
## $df
## df
## 3

## $V
## [1] 0.1809159
##
## $p_value
## [1] 0
##
## $chi2
## [1] 11910.26
##
## $df
## df
## 5
```

```

## $V
## [1] 0.01177973
##
## $p_value
## [1] 1.143897e-06
##
## $chi2
## [1] 50.49384
##
## $df
## df
## 12

```



```

## # A tibble: 10 x 2
##   cod_dept_cat     n
##   <chr>        <int>
## 1 1             67166
## 2 16            42029
## 3 10            23400
## 4 12            21688
## 5 4              20199
## 6 9              19815
## 7 3              17571
## 8 11             17530
## 9 22             13385
## 10 6             13318

```

```

## # A tibble: 10 x 2
##   cod_dept_cat     n
##   <chr>        <int>
## 1 1             26530
## 2 9              5437
## 3 5              3481
## 4 22             3064
## 5 12             3041
## 6 10             2594
## 7 13             2480
## 8 18             2313
## 9 11             2072
## 10 6             2051

```

En las variables categóricas de violencia intrafamiliar se observa un patrón muy marcado: la mayoría de víctimas son mujeres, y en estado civil predominan casado(a) y unido(a), con una proporción no despreciable de “Ignorado” (dato faltante codificado). En la relación víctima-agresor, los casos se concentran principalmente en esposo/conviviente y ex-esposo/ex-pareja, seguidos por “otro/no especificado” y “padre/madre”, mientras que “hijo(a)” y “hermano(a)” aparecen con frecuencias mucho menores. Por mes, la distribución es relativamente pareja a lo largo del año (con leves picos como mayo/julio/agosto) y un grupo pequeño en “Ignorado”, lo que sugiere estacionalidad débil. En divorcios, la escolaridad del hombre y de la mujer muestra un comportamiento similar: “Ignorado” es la categoría más frecuente, seguida por Primaria, luego Ninguno, y finalmente Media/Diversificado. Al evaluar asociación entre categóricas, las pruebas salen “significativas” por el tamaño muestral, pero los tamaños de efecto indican que Sexo vs Estado civil es muy débil (Cramér's V 0.063), Sexo vs Relación es la más relevante (V 0.181, asociación pequeña-moderada) y Mes vs Sexo es prácticamente nula (V 0.012). Finalmente, en el ranking por departamento, el código 1 lidera ampliamente tanto en violencia como en divorcios, lo cual sugiere que parte del patrón agregado está influido por el tamaño poblacional y el volumen de registros del departamento.

Preguntas de investigación basadas en hipótesis

En esta sección se plantean hipótesis preliminares basadas en creencias o supuestos comunes sobre la violencia intrafamiliar y los procesos de divorcio. Cada hipótesis será evaluada mediante análisis de datos, con el objetivo de confirmar o refutar dichas creencias.

Hipótesis 1: La violencia intrafamiliar afecta principalmente a mujeres jóvenes

Se plantea la hipótesis de que la mayoría de víctimas de violencia intrafamiliar son mujeres jóvenes. Esta creencia se basa en patrones sociales y reportes mediáticos que suelen asociar este fenómeno con mujeres en edad reproductiva.

Para evaluar esta hipótesis se analizará la distribución de la edad de la víctima y su relación con el sexo.

```

##
##      1      2
## 41796 322092

##
##      1      2
## 11.48595 88.51405

##   vic_sexo vic_edad
## 1           1 40.95370
## 2           2 33.31964

```

Resultado e interpretación El análisis muestra que el 88.5% de las víctimas registradas corresponden a mujeres, mientras que solo el 11.5% corresponde a hombres. Esta diferencia evidencia una marcada concentración del fenómeno en mujeres.

Asimismo, la edad promedio de las mujeres víctimas es de aproximadamente 33 años, con una mediana de 31 años, lo cual indica que la mayoría de los casos se concentra en adultos jóvenes.

En consecuencia, la hipótesis planteada se confirma: la violencia intrafamiliar afecta predominantemente a mujeres jóvenes dentro del conjunto de datos analizado.

Hipótesis 2: Los hombres son significativamente mayores que las mujeres en los divorcios

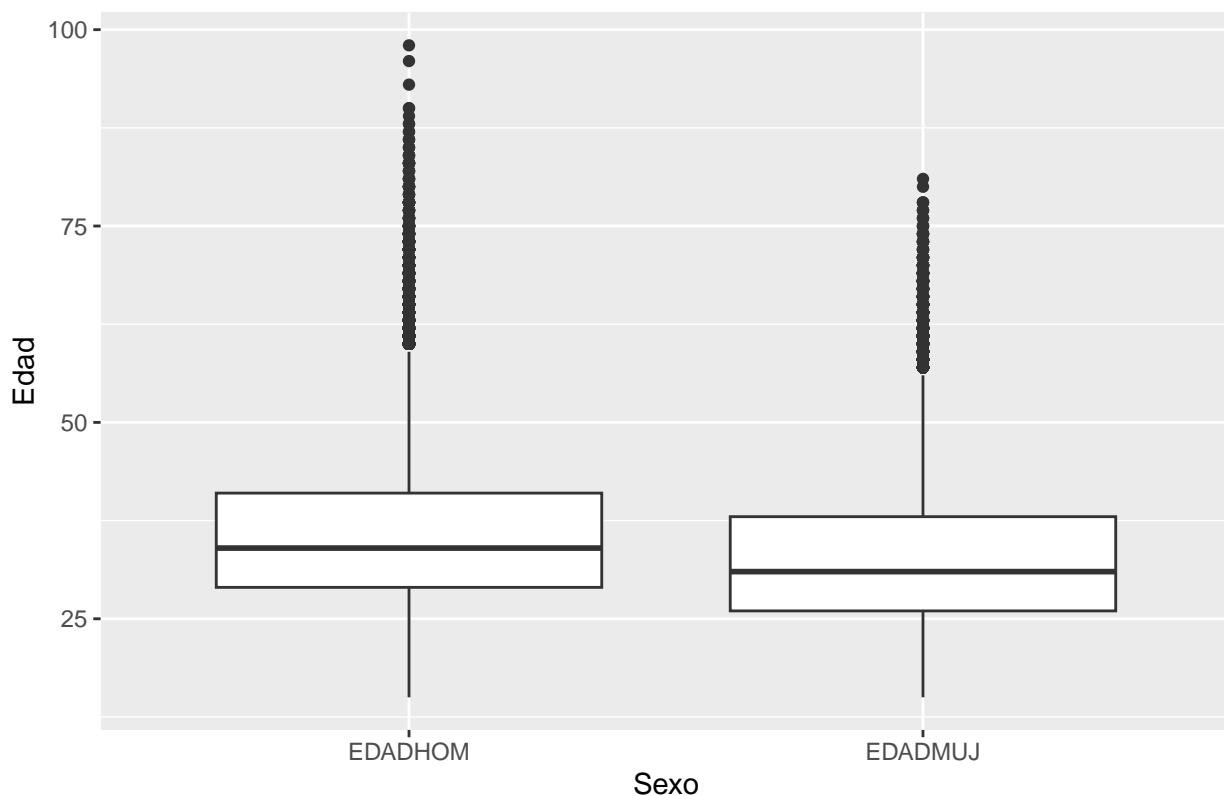
Existe la creencia de que los hombres tienden a ser mayores que las mujeres al momento del divorcio. Para evaluar esta hipótesis se compararán las edades promedio y medianas de ambos grupos.

```
## [1] 2.654367
```

```
##      Min.   1st Qu.   Median   Mean   3rd Qu.   Max. 
## -980.000    0.000    0.000   2.654    3.000  983.000
```

```
##      Min.   1st Qu.   Median   Mean   3rd Qu.   Max. 
## -41.000   0.000    2.000   3.163   6.000  75.000
```

Comparación de edad en procesos de divorcio



Resultado e interpretación Se calculó la diferencia entre la edad del hombre y la edad de la mujer en cada proceso de divorcio. La mediana de la diferencia es de 2 años, lo que indica que, en la mayoría de los casos, el hombre es ligeramente mayor que la mujer.

El tercer cuartil muestra que en el 75% de los casos el hombre es hasta 6 años mayor. Sin embargo, también se observan valores negativos, lo que indica que en un porcentaje significativo de casos la mujer es mayor que el hombre.

En consecuencia, la hipótesis se confirma parcialmente: los hombres tienden a ser mayores que las mujeres en los divorcios, aunque la diferencia promedio es moderada y no absoluta.

Los departamentos con mayor violencia tambien representan más divorcios

Si un departamento tiene muchos casos de violencia intrafamiliar -> ¿también tiene muchos divorcios?

```
## # A tibble: 6 x 2
##   depto_mcpio casos_violencia
##       <dbl>           <int>
## 1          101        29330
## 2          102         795
## 3          103         424
## 4          104         109
## 5          105         673
## 6          106        3122

## # A tibble: 6 x 2
##   DEPREG casos_divorcio
##       <dbl>           <int>
## 1          1        27573
## 2          2         1333
## 3          3         1534
## 4          4         1592
## 5          5         3187
## 6          6         1981

## # A tibble: 6 x 2
##   depto_mcpio casos_violencia
##       <dbl>           <int>
## 1          101        29330
## 2          102         795
## 3          103         424
## 4          104         109
## 5          105         673
## 6          106        3122

## # A tibble: 6 x 2
##   departamento casos_violencia
##       <dbl>           <int>
## 1             1        68182
## 2             2        11039
## 3             3        17645
## 4             4        20212
## 5             5         9965
## 6             6        13357
```

```

## # A tibble: 22 x 3
##   departamento  casosViolencia  casosDivorcio
##       <dbl>          <int>          <int>
## 1 1              68182         27573
## 2 2              11039         1333
## 3 3              17645         1534
## 4 4              20212         1592
## 5 5              9965          3187
## 6 6             13357         1981
## 7 7              6915          796
## 8 8              5484          1152
## 9 9              19929         5771
## 10 10             23801         2424
## 11 11             17580         2116
## 12 12             21515         2766
## 13 13             13090         2381
## 14 14             8728          1918
## 15 15             10729         1213
## 16 16             42171         1564
## 17 17             11279         2020
## 18 18             8018          2255
## 19 19             8288          1749
## 20 20             5271          1800
## 21 21             7085          1822
## 22 22            13605         2629

## [1] 0.8204641

##
## Pearson's product-moment correlation
##
## data: datos_dep$casosViolencia and datos_dep$casosDivorcio
## t = 6.4181, df = 20, p-value = 2.925e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.6097915 0.9228465
## sample estimates:
## cor
## 0.8204641

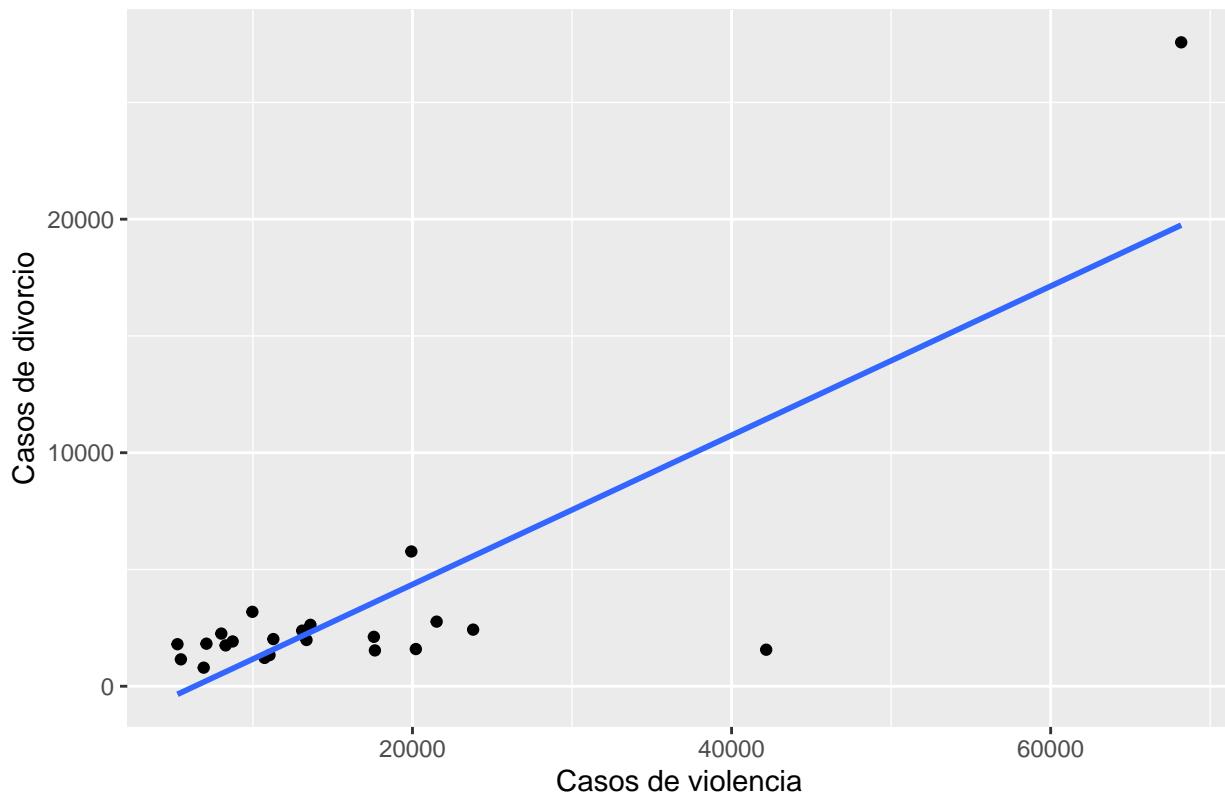
```

Resultado e interpretación Se calculó el coeficiente de correlación de Pearson entre el número de casos de violencia intrafamiliar y el número de divorcios por departamento. El resultado fue $r = 0.82$, lo que indica una correlación positiva fuerte.

El valor p obtenido ($p < 0.001$) permite rechazar la hipótesis nula de ausencia de correlación, por lo que la asociación observada es estadísticamente significativa.

En consecuencia, la hipótesis planteada se confirma: los departamentos con mayor número de casos de violencia tienden a presentar también un mayor número de divorcios. Sin embargo, es importante señalar que esta relación puede estar influenciada por el tamaño poblacional de cada departamento, por lo que no se puede establecer causalidad directa.

Relación entre violencia intrafamiliar y divorcios por departamento



El gráfico de dispersión muestra una relación positiva clara entre el número de casos de violencia intrafamiliar y el número de divorcios por departamento. La línea de tendencia ascendente indica que, en general, los departamentos con mayor cantidad de violencia también presentan un mayor número de divorcios.

Sin embargo, se observa que un departamento con valores particularmente altos influye considerablemente en la tendencia general. Esto sugiere que el tamaño poblacional puede estar afectando la relación observada. Por lo tanto, aunque la correlación es fuerte y estadísticamente significativa, no puede interpretarse como una relación causal directa.

Cabe aclarar que se pudo haber hecho el análisis por cada 100,000 habitantes pero por falta de información de habitantes por departamento.

Hipótesis 4: La mayoría de víctimas de violencia intrafamiliar tiene hijos

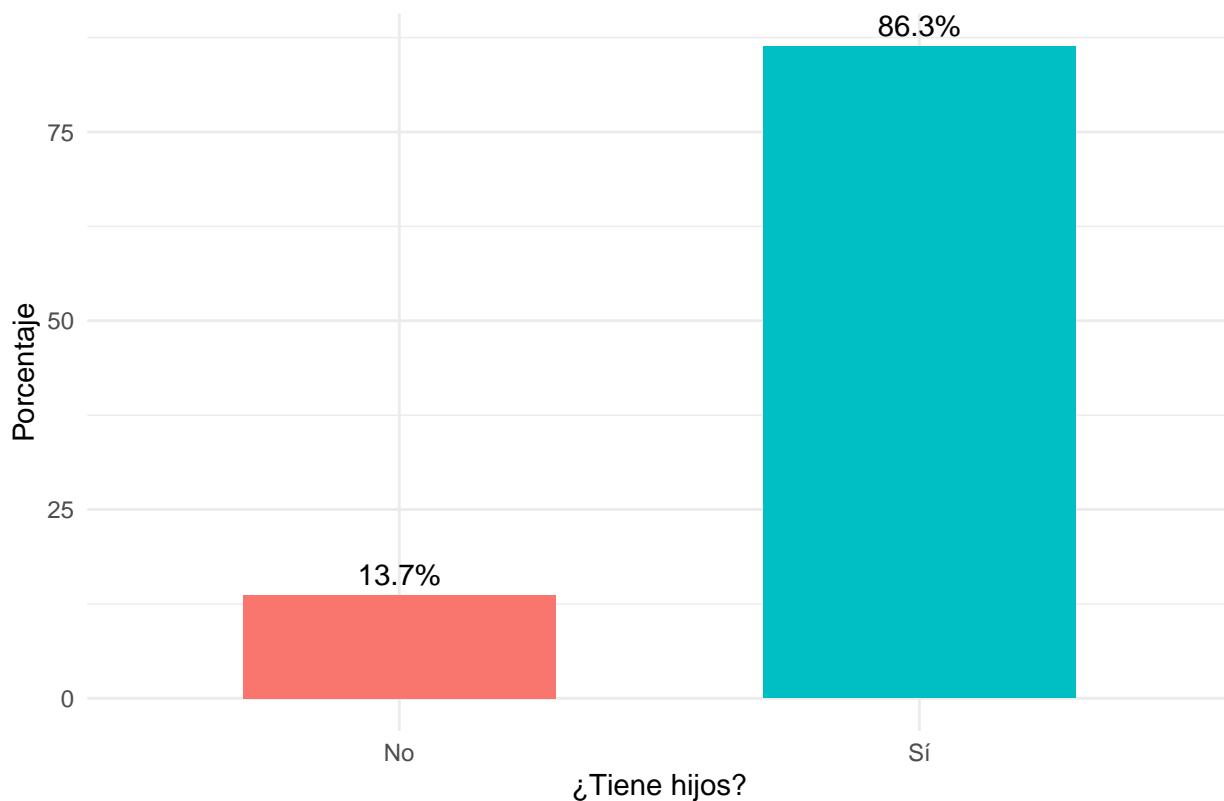
```
## # A tibble: 2 x 3
##   tiene_hijos     n  porcentaje
##   <chr>       <int>      <dbl>
## 1 No           40096     13.7
## 2 Sí          253426    86.3
```

Resultado e interpretación El análisis muestra que el 86.34% de las víctimas registradas tiene al menos un hijo, mientras que únicamente el 13.66% no tiene hijos.

En consecuencia, la hipótesis se confirma claramente: la mayoría de los casos de violencia intrafamiliar ocurre en contextos donde existen hijos dentro del entorno familiar.

Este hallazgo sugiere que el impacto del fenómeno no se limita a la víctima directa, sino que potencialmente afecta también a menores presentes en el hogar, lo cual refuerza la dimensión social y familiar del problema.

Proporción de víctimas con hijos



La gráfica refuerza visualmente que la mayoría de las víctimas de violencia intrafamiliar tiene hijos, lo cual confirma la hipótesis planteada y evidencia la dimensión familiar del fenómeno.

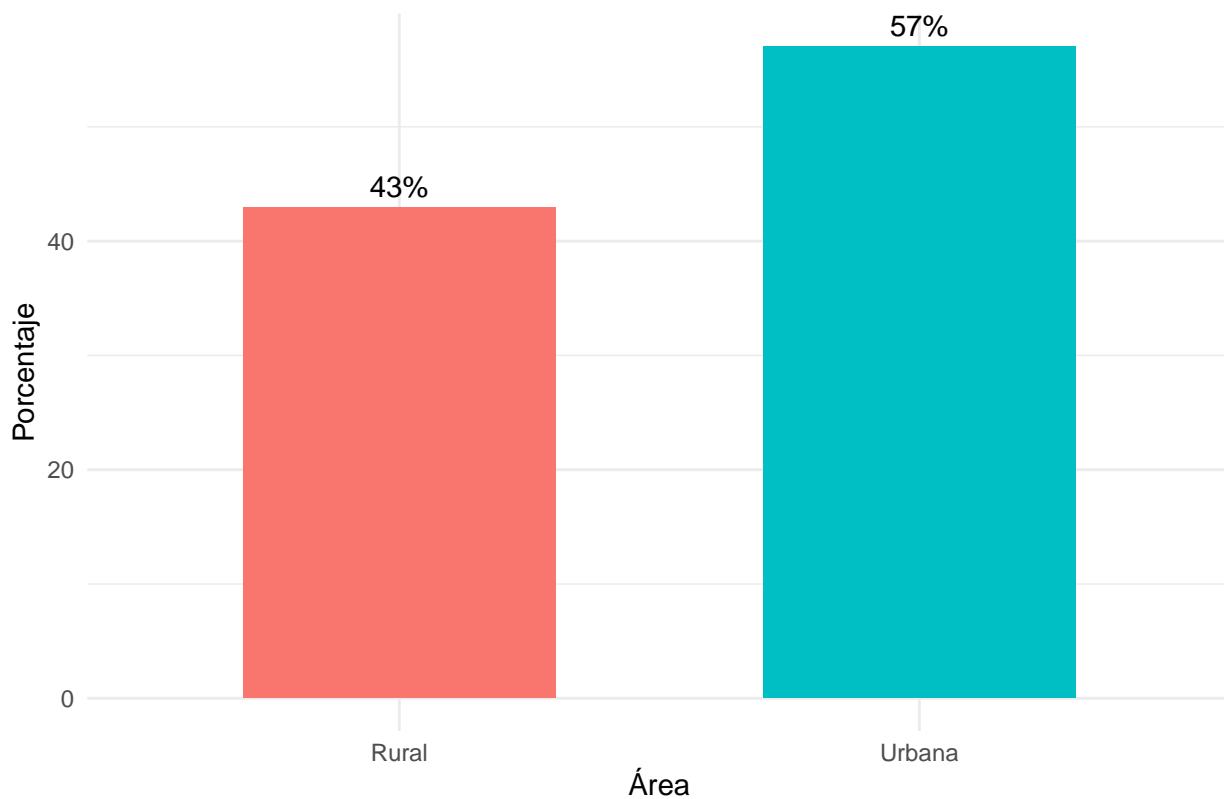
Hipótesis 5: La violencia intrafamiliar ocurre principalmente en áreas urbanas

```
##  
##      1      2      9  
## 200403 150905 12580  
  
## # A tibble: 2 x 3  
##   area     n  porcentaje  
##   <chr>   <int>     <dbl>  
## 1 Rural   150905    43.0  
## 2 Urbana  200403    57.0
```

Resultado e interpretación El análisis muestra que el 57.04% de los casos de violencia intrafamiliar ocurrió en áreas urbanas, mientras que el 42.96% se registró en áreas rurales.

En consecuencia, la hipótesis se confirma: la violencia intrafamiliar ocurre mayoritariamente en zonas urbanas. Sin embargo, la diferencia no es extrema, lo que indica que el fenómeno también presenta una presencia significativa en áreas rurales, evidenciando que se trata de un problema generalizado y no exclusivamente urbano.

Distribución de violencia por área



Aunque la mayoría de los casos se registran en áreas urbanas, esta diferencia podría estar influenciada por la mayor concentración poblacional en dichas zonas. Dado que el análisis se basa en frecuencias absolutas y no en tasas ajustadas por población, no es posible concluir que el riesgo de violencia sea mayor en el área urbana, sino únicamente que el número total de casos es superior.

Determinación del número óptimo de clusters

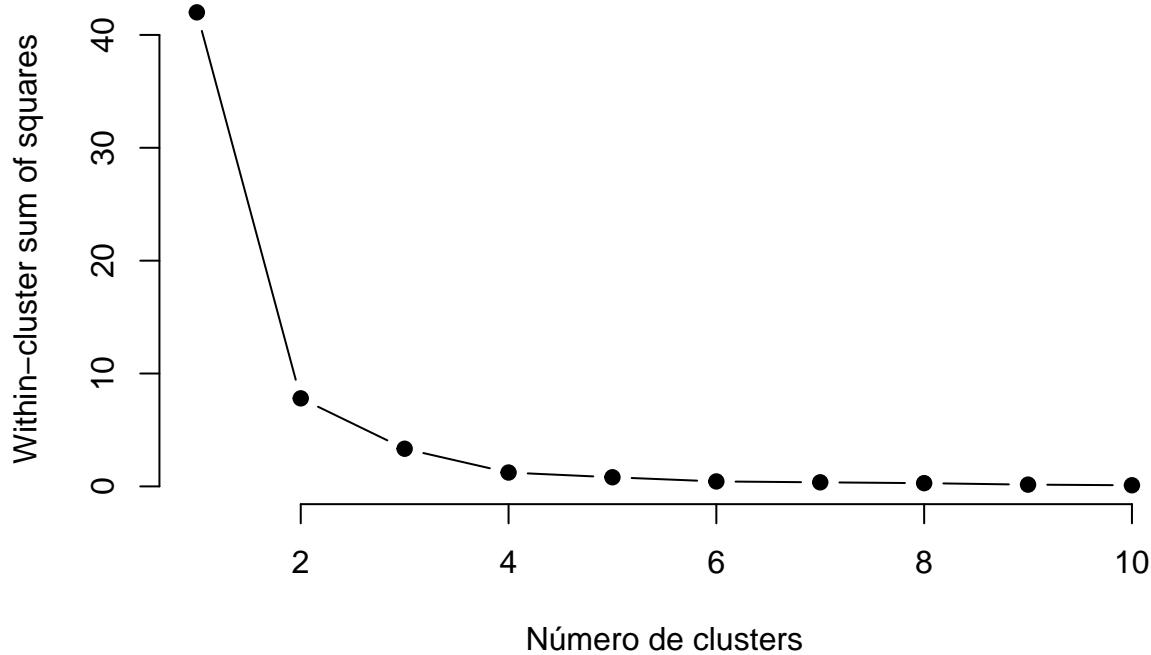
Para determinar el número adecuado de grupos se utilizaron dos métodos: el método del codo y el método de la silueta.

El método del codo sugiere un punto de inflexión alrededor de tres clusters, ya que a partir de ese valor la reducción en la variabilidad interna se vuelve menos significativa.

Por otro lado, el método de la silueta muestra su valor máximo en $k = 2$, con una anchura promedio cercana a 0.85, lo cual indica una separación muy clara entre los grupos.

Dado que la silueta evalúa directamente la cohesión y separación de los clusters, se selecciona $k = 2$ como el número óptimo de grupos para este análisis.

Método del Codo

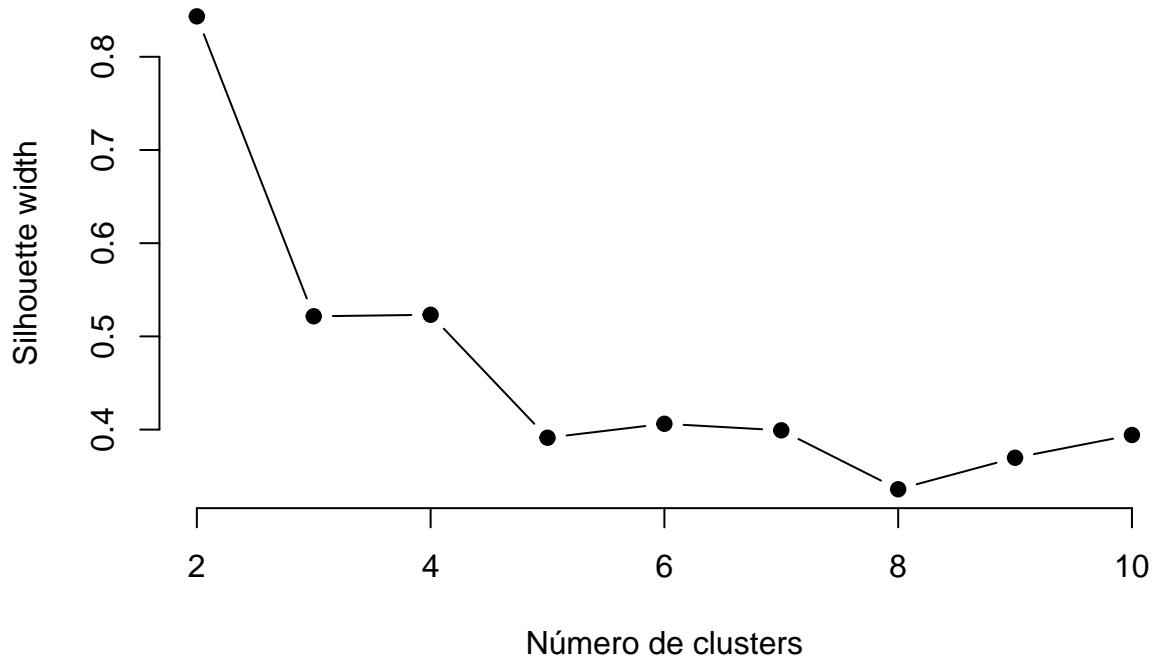


Método del Codo

El método del codo permite identificar el número adecuado de clusters analizando la variación interna (within-cluster sum of squares) a medida que aumenta el número de grupos. En la gráfica se observa una disminución pronunciada de la variabilidad interna entre uno y tres clusters. A partir de este punto, la reducción se vuelve progresivamente menor, indicando que agregar más grupos no aporta una mejora sustancial en la cohesión interna.

Este comportamiento sugiere la presencia de una estructura natural en los datos alrededor de tres clusters, donde se produce el cambio más notable en la pendiente de la curva.

Método de la Silueta



Método de la Silueta

El método de la silueta evalúa la calidad del agrupamiento midiendo qué tan bien se separan los elementos de distintos clusters y qué tan cohesionados están dentro de su propio grupo. En la gráfica se observa que el valor máximo de la anchura promedio de la silueta se alcanza cuando $k = 2$, con un valor cercano a 0.85.

Un valor de silueta superior a 0.5 indica una separación clara entre grupos, mientras que valores cercanos a 1 representan una estructura muy bien definida. Por lo tanto, el resultado sugiere que la partición en dos clusters proporciona la mejor separación y cohesión para los datos analizados.

```
## # A tibble: 22 x 4
##   departamento casosViolencia casosDivorcio cluster
##   <dbl>           <int>          <int> <fct>
## 1 1                 1            68182    27573 1
## 2 2                 2            11039     1333  2
## 3 3                 3            17645     1534  2
## 4 4                 4            20212     1592  2
## 5 5                 5            9965      3187  2
## 6 6                 6            13357     1981  2
## 7 7                 7            6915      796   2
## 8 8                 8            5484      1152  2
## 9 9                 9            19929     5771  2
## 10 10              10            23801     2424  2
## # i 12 more rows
## # A tibble: 2 x 4
##   cluster promedioViolencia promedioDivorcio n_departamentos
```

```

##   <fct>      <dbl>      <dbl>      <int>
## 1 1          68182     27573       1
## 2 2          14081.    2095.      21

```

Aplicación del algoritmo K-means

Se aplicó el algoritmo K-means con $k = 2$, determinado previamente mediante el método de la silueta. El resultado muestra que uno de los clusters contiene únicamente un departamento, caracterizado por niveles significativamente más altos de violencia y divorcios en comparación con el resto.

Este departamento corresponde al departamento de Guatemala, el cual presenta valores extremos que lo diferencian claramente del resto del país. El segundo cluster agrupa los otros 21 departamentos, que presentan niveles considerablemente menores en ambas variables.

El resultado indica que la estructura natural de los datos está fuertemente influenciada por la magnitud de los valores del departamento de Guatemala, lo cual genera una separación clara entre este y el resto.

Hallazgos del Análisis Exploratorio

- Perfil de la Víctima:** La violencia intrafamiliar posee un sesgo de género y edad muy marcado. El **88.5% de las víctimas son mujeres**, con una mediana de edad de **31 años**, confirmando que el fenómeno afecta desproporcionadamente a mujeres adultas jóvenes en etapa reproductiva y de crianza (el 86.3% tiene hijos).
- La Trampa de la Correlación:** Existe una correlación lineal fuerte ($r = 0.82$) entre la cantidad de denuncias de violencia y la cantidad de divorcios por departamento. Sin embargo, este dato es engañoso: está impulsado casi exclusivamente por el **Departamento de Guatemala**, cuyo volumen poblacional infla las cifras, actuando como un *outlier* que distorsiona la tendencia nacional.
- Homogamia en Rupturas:** En los divorcios, se rompe el mito de las grandes diferencias de edad. La mayoría de las parejas que disuelven su vínculo tienen edades similares (diferencia mediana de 2 años), lo que sugiere que las tensiones que llevan al divorcio y/o violencia ocurren entre pares generacionales.

Caracterización de los Grupos (Nombres Propuestos)

El algoritmo *K-Means* ($k = 2$) identificó una estructura polarizada en el país. Basado en sus características, se proponen los siguientes nombres para los grupos:

Grupo 1: “El Núcleo Metropolitano (Outlier)”

- Integrantes:** Únicamente el **Departamento de Guatemala**.
- Características:** Representa un escenario de “alta intensidad” que no es comparable con el resto del país. Sus cifras de violencia (aprox. 68k casos) y divorcios (27k) rompen la escala, siendo entre 6 y 20 veces superiores al promedio de los demás departamentos. Estadísticamente, es un caso atípico extremo.

Grupo 2: “El Interior Estándar”

- Integrantes:** Los **21 departamentos restantes** (Huehuetenango, San Marcos, Petén, Zacapa, etc.).
- Características:** A pesar de sus diferencias culturales y geográficas, estos departamentos se agrupan por tener volúmenes de casos “bajos” o “medios” en comparación con la capital. Para el algoritmo, todos son similares simplemente porque “no son la capital”.

Plan de Siguientes Pasos

Para corregir el sesgo identificado y profundizar en el análisis, se propone la siguiente hoja de ruta:

1. **Zoom-In (Sub-segmentación):** Ejecutar una segunda iteración de *clustering excluyendo al Departamento de Guatemala*. Al retirar este *outlier*, se podrán revelar los matices y diferencias reales entre los departamentos del interior (ej. diferenciar zonas de alta violencia en oriente vs. occidente).
2. **Normalización Demográfica:** Sustituir el análisis de frecuencias absolutas por **Tasas por cada 100,000 habitantes**. Esto es crucial para determinar si departamentos pequeños (como El Progreso o Zacapa) tienen en realidad una *intensidad* de violencia mayor que la capital, aunque tengan menos casos totales.
3. **Variable de Ratio:** Crear y analizar la métrica *Divorcios / Denuncias de Violencia*. Esto permitirá identificar departamentos donde existe una alta denuncia pero baja disolución legal, lo cual podría indicar barreras culturales, económicas o legales específicas de esa región.