Proyecto 1

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

2026-02-13

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 existan 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 decada 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 - ADRIANA ADRIANA ADRIANA

Enuncia el problema científico que se desprende de la situación planteada. Comprende bien cuál es el problema. La idea es que en base a la idea planteada anteriormente, se pueda justificar con algún artículo la presencia de patrones peculiares en el matrimonio/divorcio en los últimos años, o que hablen acerca de las percepciones sociales. Incluir 2 citas como mínimo, citar y colocar referencias en APA-7 al final. La idea es justificar con datos puntuales, por qué la investigación que estamos realizando es relevante.

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()
}

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)
```

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:

```r
dim(divorcios_depto)
```

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

Se tienen 4186 observaciones y 5 variables.

Divorcios por rangos de edad:

```r
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, habían 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.

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

Análisis Exploratorio

Exploración de Variables Númericas PEDRITO PEDRITO PEDRITO

(Medidas de tendencia central, distribución y orden. (todas las variables tienen que tener una análisis sobre los resultados, se debe incluir histogramas, boxplots o gráficos de dispersión, uno como mínimo y explicarlo). Para cada una hay que describir su distribución (sacarle su p-value) Si o si piden analizar las variables por tecnicas de estadistica descriptiva, tener graficos, y sacar la distribucion.)

```r
print("Matrimonios por deparatmento")
```

Medidas de tendencia central

```
## [1] "Matrimonios por deparatmento"
```

```r
summary(matrimonios_depto)
```

```
##       anio       nivel_geo         departamento            mes
##  Min.   :2009   Length:4186        Length:4186        Min.   : 1.00
##  1st Qu.:2012   Class :character   Class :character   1st Qu.: 3.75
##  Median :2016   Mode  :character   Mode  :character   Median : 6.50
##  Mean   :2016                                         Mean   : 6.50
##  3rd Qu.:2019                                         3rd Qu.: 9.25
##  Max.   :2022                                         Max.   :12.00
##                                                       NA's   :322
##      valor
##  Min.   :    8
##  1st Qu.:  144
##  Median :  235
##  Mean   : 1009
##  3rd Qu.:  437
##  Max.   :87480
##
```

```r
print(" ")
```

```
## [1] " "
```

```r
print("Matrimonios por edad")
```

```
## [1] "Matrimonios por edad"
```

```r
summary(matrimonios_edad)
```

```
##       anio      edad_mujer_grupo   edad_hombre_grupo      valor
##  Min.   :2009   Length:2191        Length:2191        Min.   :    0.0
##  1st Qu.:2012   Class :character   Class :character   1st Qu.:    1.0
##  Median :2015   Mode  :character   Mode  :character   Median :   31.0
##  Mean   :2015                                         Mean   :  481.9
##  3rd Qu.:2019                                         3rd Qu.:  217.0
##  Max.   :2022                                         Max.   :14859.0
```

```r
print(" ")
```

```
## [1] " "
```

```r
print("Divorcios por departamento")
```

```
## [1] "Divorcios por departamento"
```

```r
summary(divorcios_depto)
```

```
##       anio       nivel_geo         departamento            mes        
##  Min.   :2009   Length:4186        Length:4186        Min.   : 1.00  
##  1st Qu.:2012   Class :character   Class :character   1st Qu.: 3.75  
##  Median :2016   Mode  :character   Mode  :character   Median : 6.50  
##  Mean   :2016                                         Mean   : 6.50  
##  3rd Qu.:2019                                         3rd Qu.: 9.25  
##  Max.   :2022                                         Max.   :12.00  
##                                                       NA's   :322    
##      valor        
##  Min.   :   0.0   
##  1st Qu.:   8.0   
##  Median :  14.0   
##  Mean   :  78.9   
##  3rd Qu.:  25.0   
##  Max.   :9950.0   
##
```

```r
print(" ")
```

```
## [1] " "
```

```r
print("Divorcios por edad")
```

```
## [1] "Divorcios por edad"
```

```r
summary(divorcios_edad)
```

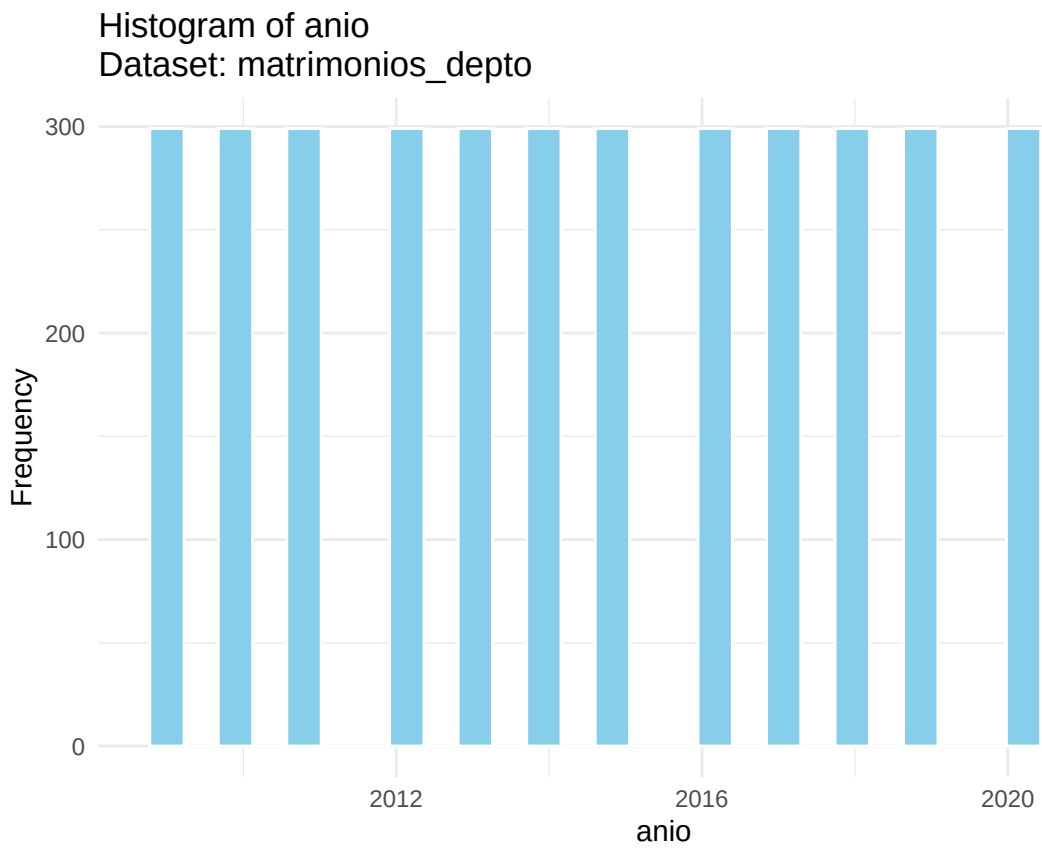
```
##       anio      edad_mujer_grupo   edad_hombre_grupo      valor          
##  Min.   :2009   Length:1878        Length:1878        Min.   :   0.00  
##  1st Qu.:2012   Class :character   Class :character   1st Qu.:   0.00  
##  Median :2015   Mode  :character   Mode  :character   Median :   4.00  
##  Mean   :2015                                         Mean   :  43.97  
##  3rd Qu.:2019                                         3rd Qu.:  21.00  
##  Max.   :2022                                         Max.   :3582.00
```

```r
print(" ")
```

```
## [1] " "
```

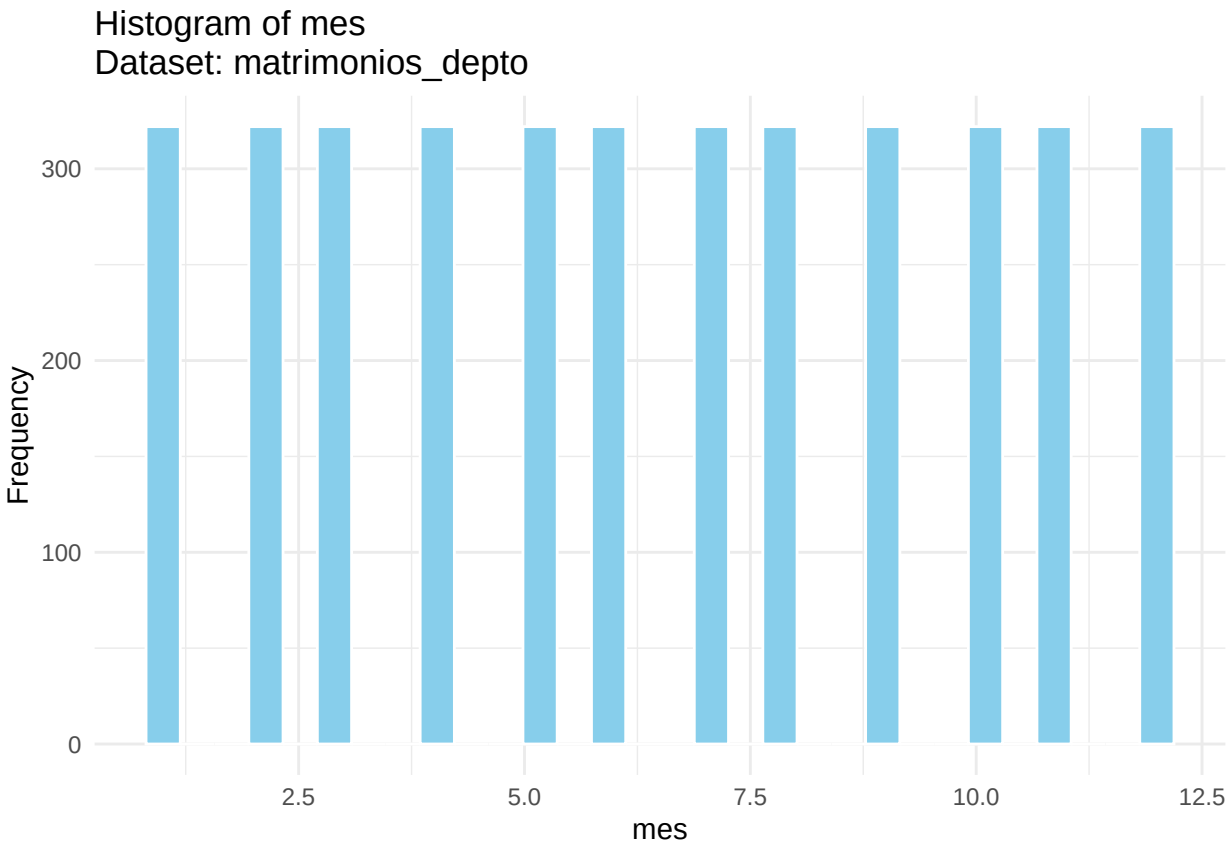
Del summary ya podemos sacar unos cuantos datos cuantitativos los cuales no pueden servir probablmente un poco a futuro, como que la media de divorcios por anio y matrimonio por anio tienen un anio de diferencia

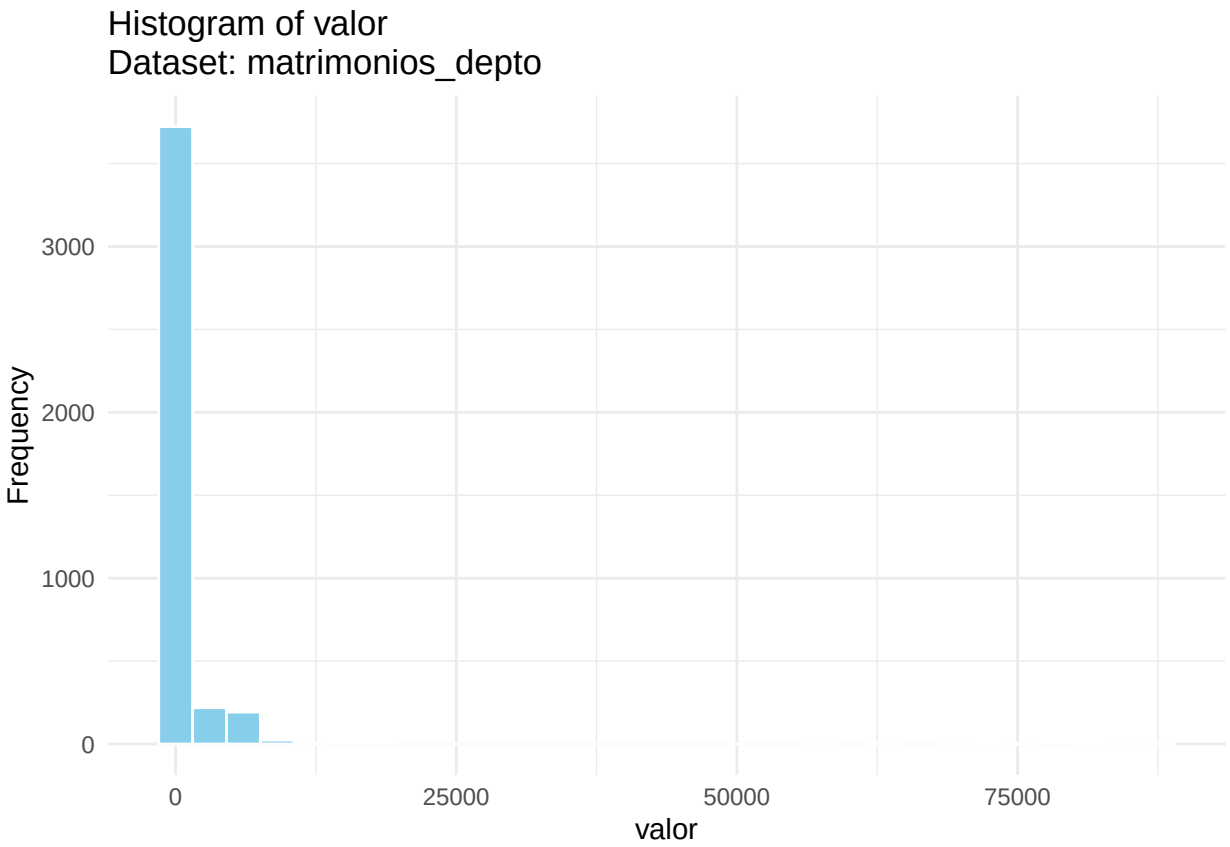
```r
each_df <- list(
  matrimonios_depto = matrimonios_depto,
  matrimonios_edad = matrimonios_edad,
  divorcios_depto = divorcios_depto,
  divorcios_edad = divorcios_edad
)
plot_histograms_gg <- function(df, df_name, bins = 30) {

  df <- as.data.frame(df)
  numeric_vars <- df[, sapply(df, is.numeric), drop = FALSE]

  for (col_name in names(numeric_vars)) {

    p <- ggplot(df, aes(x = .data[[col_name]])) +
      geom_histogram(bins = bins, fill = "skyblue", color = "white") +
      labs(
        title = paste("Histogram of", col_name, "\nDataset:", df_name),
        x = col_name,
        y = "Frequency"
      ) +
      theme_minimal()

    print(p)
  }
}

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

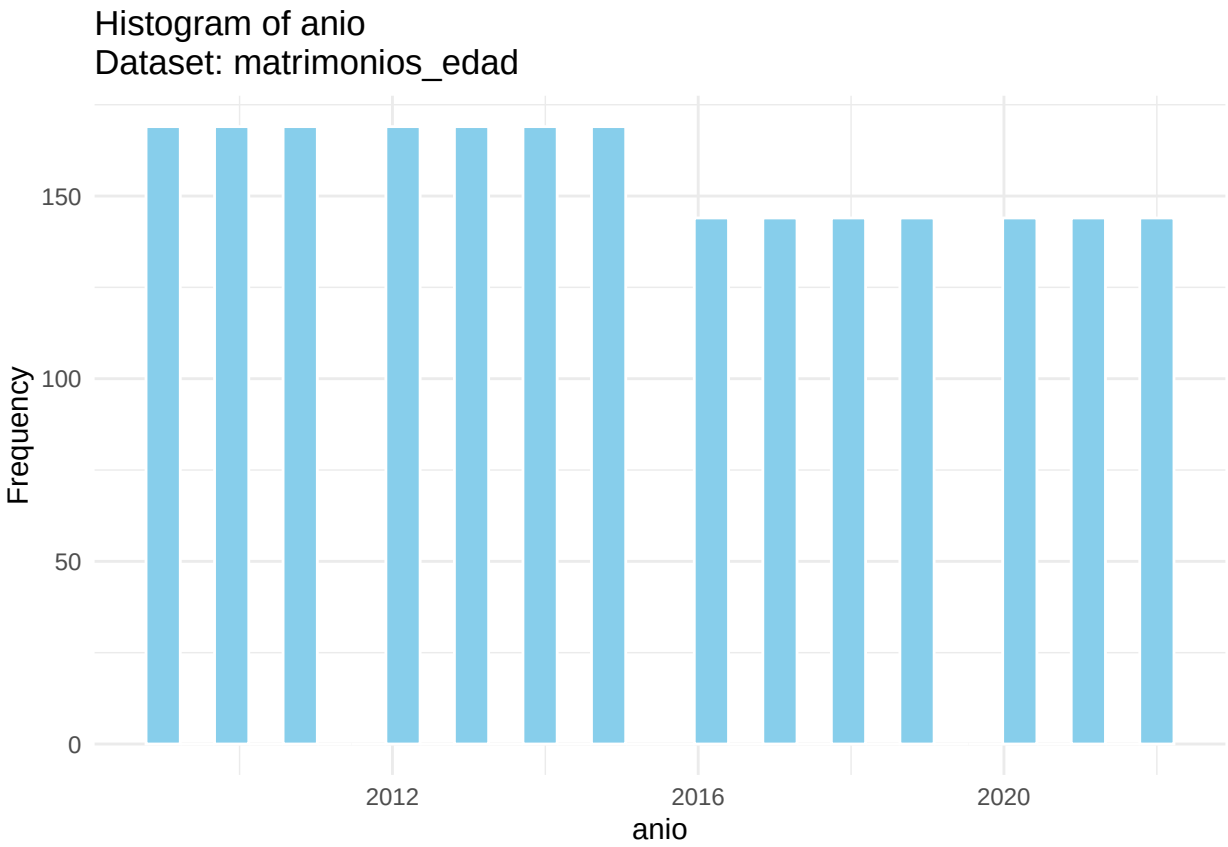


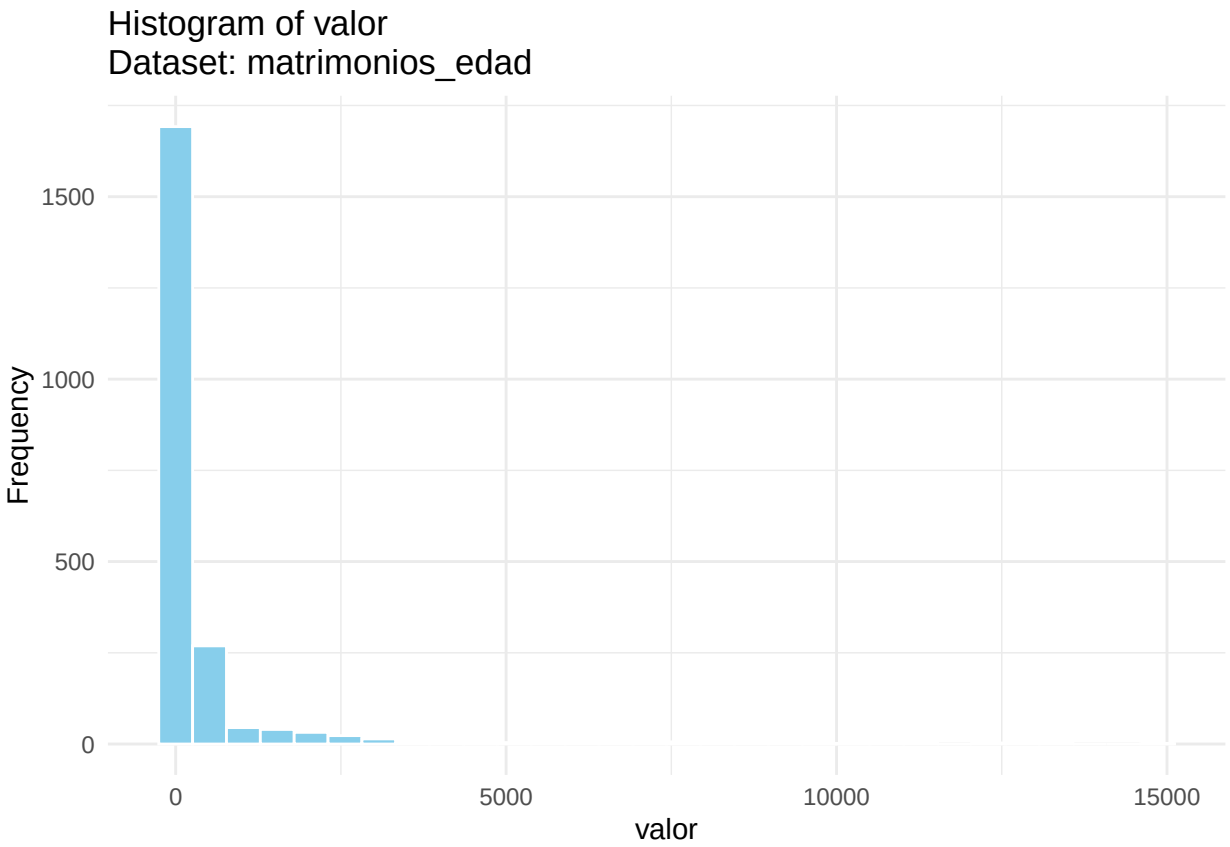
Distribuciones por variables

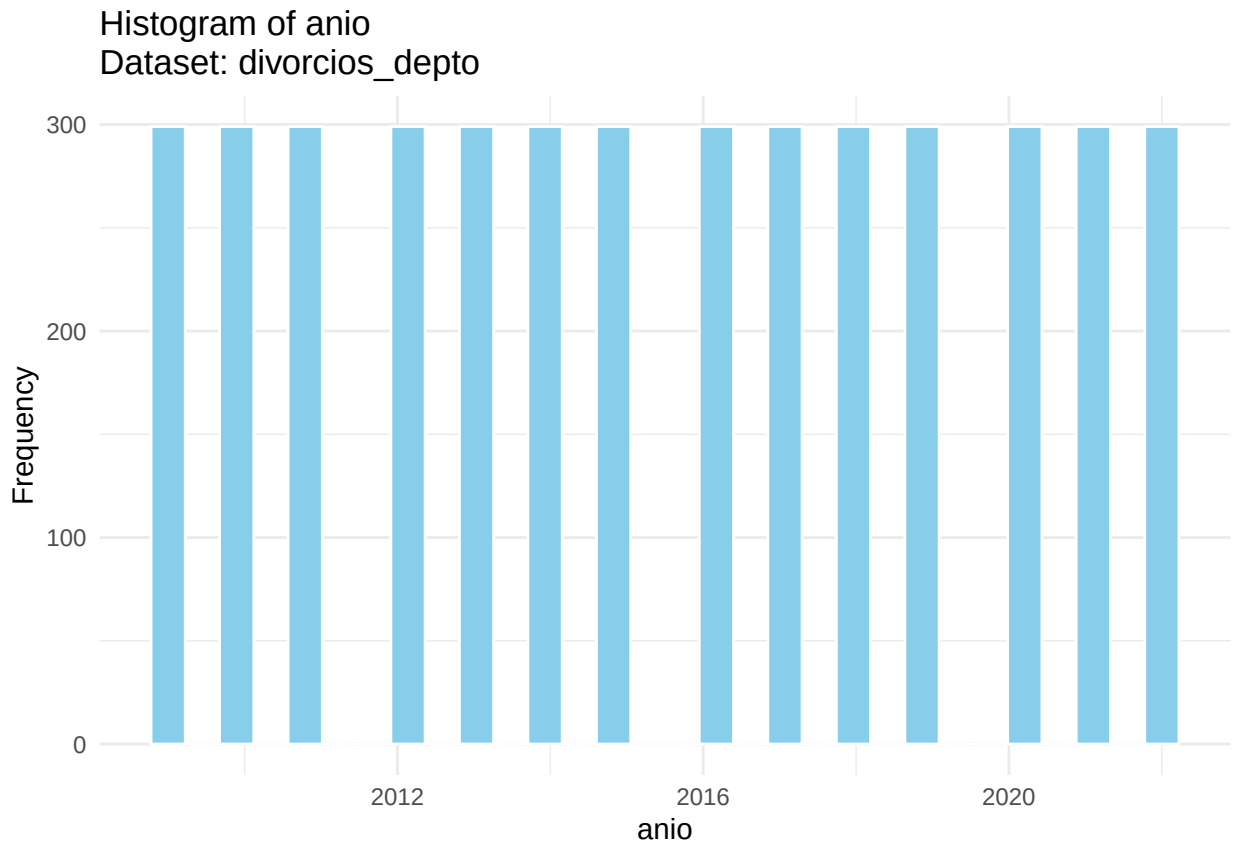
```
## Warning: Removed 322 rows containing non-finite outside the scale range
## (`stat_bin()`).
```


Histogram of mes
Dataset: matrimonios_depto
Frequency
300
200
100
0
2.5
5.0
7.5
10.0
12.5
mes

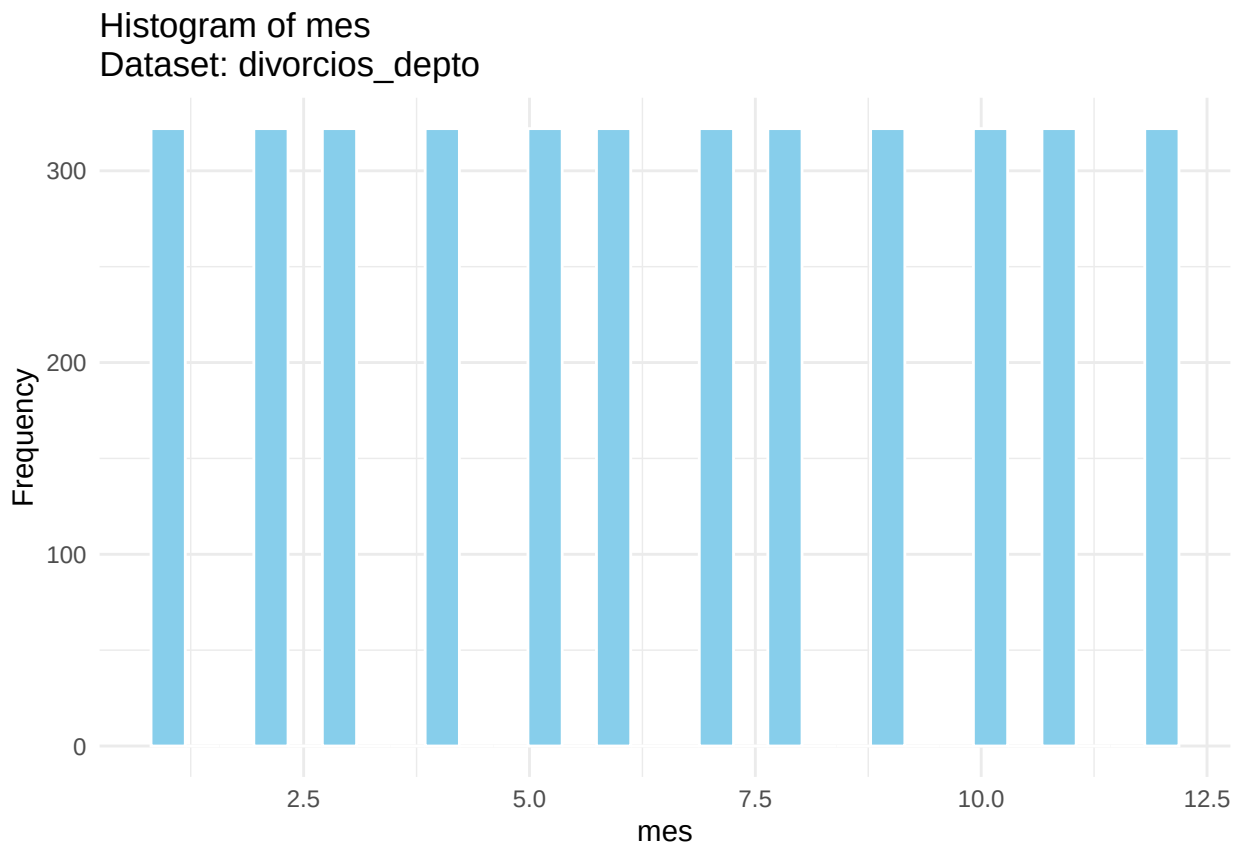

Histogram of valor
Dataset: matrimonios_depto
Frequency
3000
2000
1000
0
0
25000
50000
75000
valor

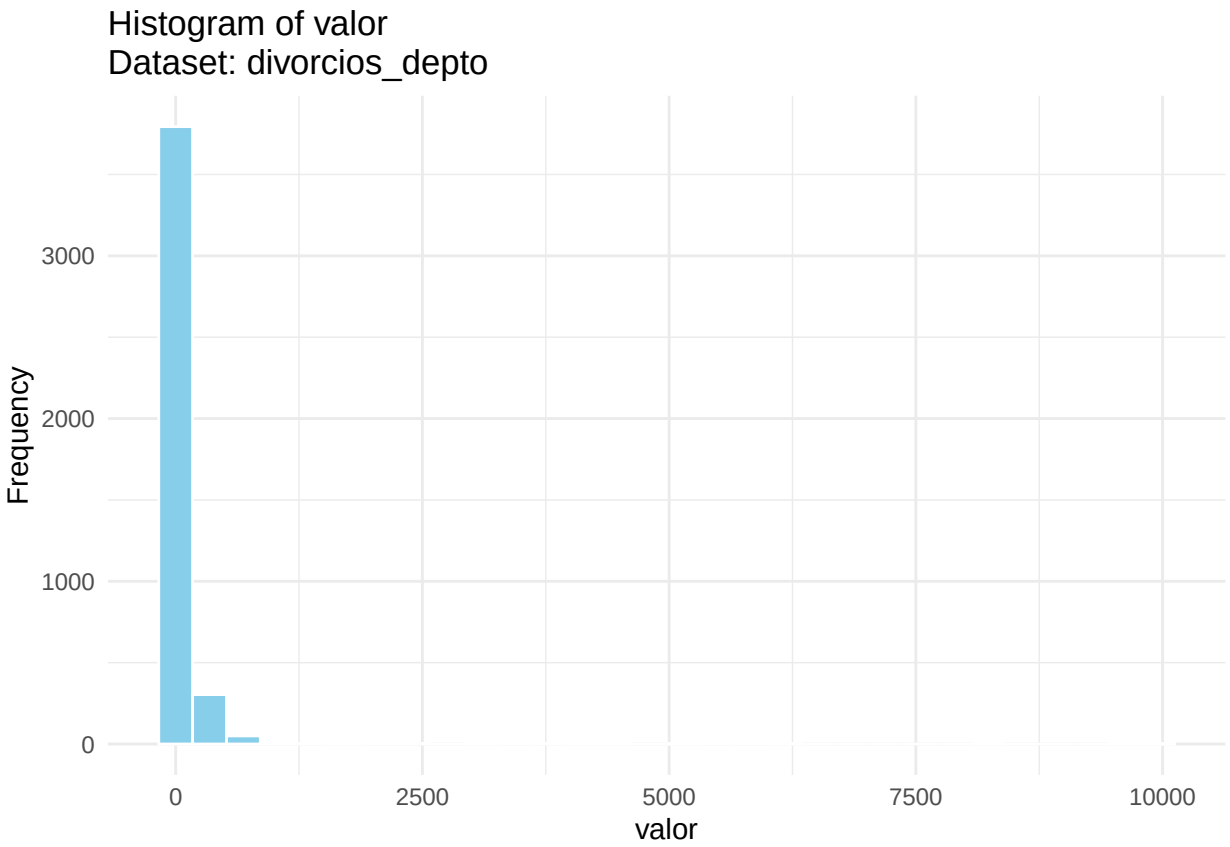

Histogram of anio
Dataset: matrimonios_edad
Frequency
150
100
50
0
2012
2016
2020
anio

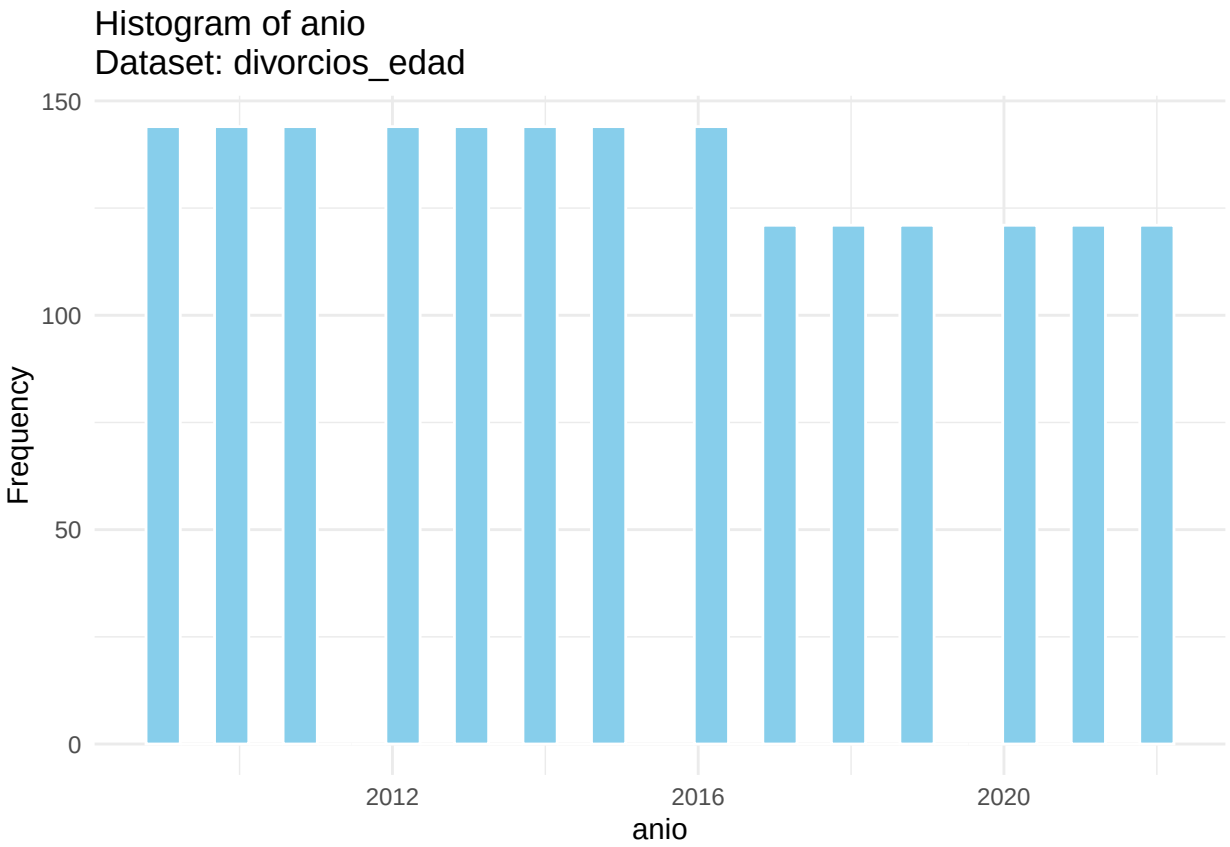

Histogram of valor
Dataset: matrimonios_edad
Frequency
1500
1000
500
0
0
5000
10000
15000
valor

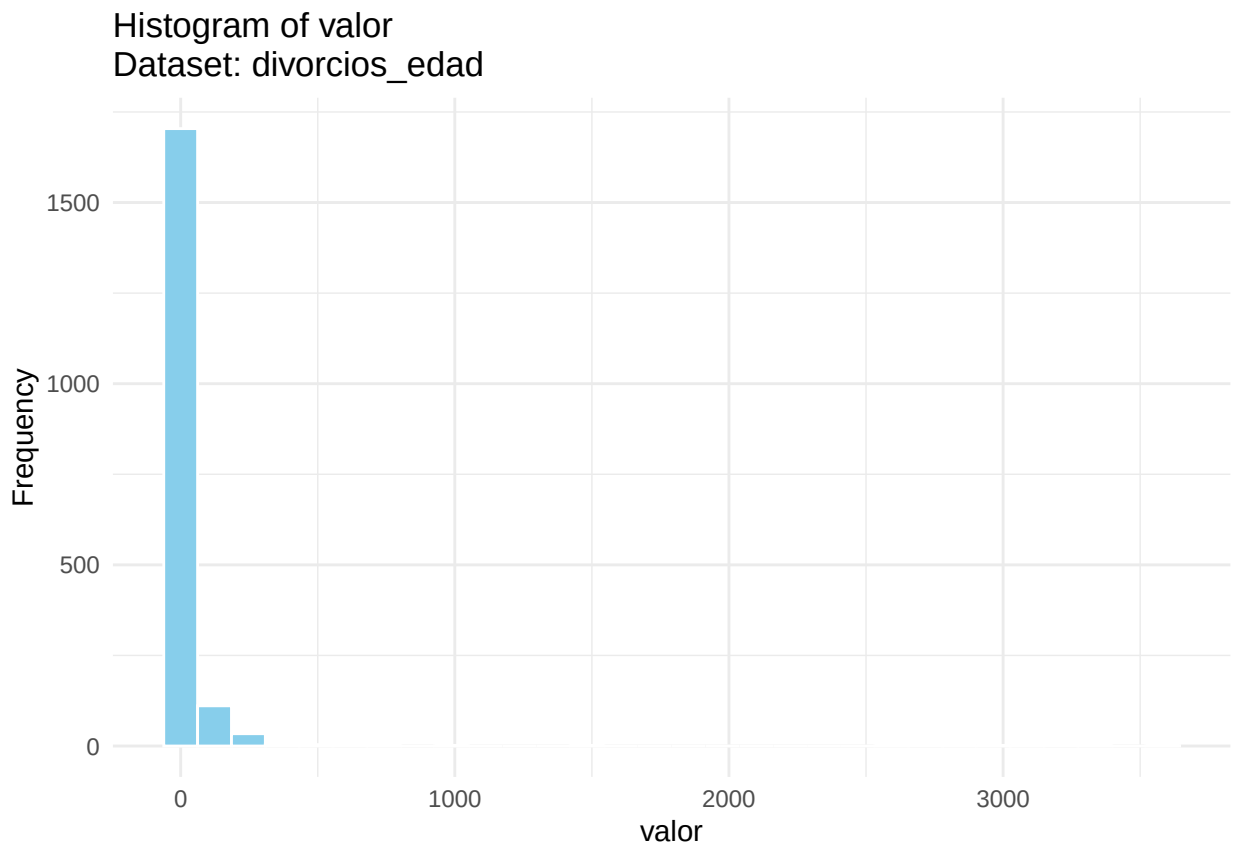



```
## Warning: Removed 322 rows containing non-finite outside the scale range
## (‘stat_bin()‘).
```


Histogram of mes
Dataset: divorcios_depto
Frequency
300
200
100
0
2.5
5.0
7.5
10.0
12.5
mes


Histogram of valor
Dataset: divorcios_depto
Frequency
3000
2000
1000
0
0
2500
5000
7500
10000
valor


Histogram of anio
Dataset: divorcios_edad
150
100
50
0
Frequency
2012
2016
2020
anio



Como podemos ver de los histogramas, muchas columnas no siguen una distribucion muy clara, como si N cantidad de muestras fueron tomadas siempre por cada mes o anio. Las que si nos importan son como tal la cantidad de divorcios y la cantidad de matrimonios, los cuales a simple vista no parecen que tengan una distribucion normal para nada, si no mas sesagadas a la derecha.

Ademas de eso pareciera al ver el histograma pareciera que tienen una clase de "corelacion" los datos, teniendo un pico extremadamente alto de cantidad de divorcios a cierto punto igual que con los matrimonios, del lado del valor.

```
prueba_lilliefors <- function(df, df_name, alpha = 0.05) {

  df <- as.data.frame(df)
  numeric_vars <- df[, sapply(df, is.numeric), drop = FALSE]

  resultados <- data.frame(
    Dataset = character(),
    Variable = character(),
    P_value = numeric(),
    Normalidad = character(),
    stringsAsFactors = FALSE
  )

  for (col_name in names(numeric_vars)) {
```

```
    x <- na.omit(numeric_vars[[col_name]])

    test <- lillie.test(x)

    decision <- ifelse(test$p.value < alpha,
                       "No tiene normalidad",
                       "No se puede determinar normalidad")

    resultados <- rbind(resultados, data.frame(
      Dataset = df_name,
      Variable = col_name,
      P_value = test$p.value,
      Normalidad = decision
    ))
  }

  return(resultados)
}

resultados_totales <- data.frame()

for (df_name in names(each_df)) {
  res <- prueba_lilliefors(each_df[[df_name]], df_name)
  resultados_totales <- rbind(resultados_totales, res)
}

print(resultados_totales)
```

Distribucion normal

```
##              Dataset Variable       P_value          Normalidad
## 1  matrimonios_depto     anio  2.239182e-98 No tiene normalidad
## 2  matrimonios_depto      mes 9.717004e-103 No tiene normalidad
## 3  matrimonios_depto    valor  0.000000e+00 No tiene normalidad
## 4   matrimonios_edad     anio  1.989263e-55 No tiene normalidad
## 5   matrimonios_edad    valor  0.000000e+00 No tiene normalidad
## 6    divorcios_depto     anio  2.239182e-98 No tiene normalidad
## 7    divorcios_depto      mes 9.717004e-103 No tiene normalidad
## 8    divorcios_depto    valor  0.000000e+00 No tiene normalidad
## 9     divorcios_edad     anio  5.925539e-46 No tiene normalidad
## 10    divorcios_edad    valor  0.000000e+00 No tiene normalidad
```

Con el alpha que pusimos podemos percartarnos que ninguno de los datos tiene una distrubucion normal, como anteriormente visto facilmente de la forma grafica

Exploración de Variables Categóricas BENITEZ BENITEZ BENITEZ

Tablas de frecuencia Gráficos de barra para cada una

Relaciones entre Variables

(este es un poquito peculiar, por como estan los datos, se me ocurre que vean si existe una relación entre la cantidad de casamientos con el rango de edad de la novia o del hombre, y luego, ver si hay una relación entre la cantidad de divorcios y los rangos de edades de los novios/novias).

Clustering

Tendencia de Agrupamiento y Número de Clusters

```
# Quedarnos solo con datos departamentales, y segmentados por datos anuales, en vez de mes x año
matrimonios_anual <- matrimonios_depto %>%
  filter(
    nivel_geo == "departamento",
    is.na(mes)
  ) %>%
  select(anio, departamento, matrimonios = valor)

divorcios_anual <- divorcios_depto %>%
  filter(
    nivel_geo == "departamento",
    is.na(mes)
  ) %>%
  select(anio, departamento, divorcios = valor)

# Unificar datos anuales de matrimonios y divorcios
matrimonios_divorcios_depto_anual <- full_join(
  matrimonios_anual,
  divorcios_anual,
  by = c("anio", "departamento")
)

# Variable para analizar el ratio entre divorcios y matrimonios
matrimonios_divorcios_depto_anual <- matrimonios_divorcios_depto_anual %>%
  mutate(
    ratio_div_matr = divorcios / matrimonios
  )

matrimonios_divorcios_depto_anual
```

```
## # A tibble: 308 x 5
##     anio departamento  matrimonios divorcios ratio_div_matr
##    <dbl> <chr>               <dbl>     <dbl>          <dbl>
##  1  2009 alta verapaz         4300       129         0.03
##  2  2009 baja verapaz         1334        66         0.0495
##  3  2009 chimaltenango        3099        56         0.0181
##  4  2009 chiquimula           1766        91         0.0515
##  5  2009 el progreso           786        47         0.0598
##  6  2009 escuintla            2859       160         0.0560
##  7  2009 guatemala           11824      1024         0.0866
##  8  2009 huehuetenango        5276        84         0.0159
##  9  2009 izabal               1276        86         0.0674
```

```
## 10  2009 jalapa               1289        79         0.0613
## # i 298 more rows
```
