EDA Salud Mental

Table of contents

0.1 Librerías

```
library(tidyverse)
```

```
Warning: package 'tidyverse' was built under R version 4.4.3
```

```
Warning: package 'ggplot2' was built under R version 4.4.3
```

```
Warning: package 'dplyr' was built under R version 4.4.3
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   4.0.0     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.4
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(readxl)
```

```
Warning: package 'readxl' was built under R version 4.4.3
```

```
library(visdat)
```

```
Warning: package 'visdat' was built under R version 4.4.3
```

```
library(GGally)
```

```
Warning: package 'GGally' was built under R version 4.4.3
```

```
library(patchwork)
```

```
Warning: package 'patchwork' was built under R version 4.4.3
```

```r
library(ggridges)
```

```
Warning: package 'ggridges' was built under R version 4.4.3
```

```r
# Opciones de gráficos para acelerar el render a PDF
knitr::opts_chunk$set(
  dev = "png",
  dpi = 200,
  fig.width = 11,
  fig.height = 6.5,
  out.width = "100%"
)
```

0.2 Carga de datos

Leemos el Excel original. Si aparecen warnings en la carga, se muestran justo encima de esta línea en la salida y los comentamos a continuación.

```r
SaludMental <- read_excel("./SaludMental.xls")
```

```
Warning: Expecting logical in CX1831 / R1831C102: got '0UT94ZZ'

Warning: Expecting logical in CY1831 / R1831C103: got '0UT24ZZ'

Warning: Expecting logical in CZ1831 / R1831C104: got '0UT74ZZ'

Warning: Expecting logical in AL5125 / R5125C38: got 'GZB4ZZZ'

Warning: Expecting logical in AL5554 / R5554C38: got 'GZB4ZZZ'

Warning: Expecting logical in AM5554 / R5554C39: got 'GZB4ZZZ'

Warning: Expecting logical in AN5554 / R5554C40: got 'GZB4ZZZ'

Warning: Expecting logical in AO5554 / R5554C41: got 'GZB4ZZZ'

Warning: Expecting logical in AP5554 / R5554C42: got 'GZB4ZZZ'
```

```
Warning: Expecting logical in AQ5554 / R5554C43: got 'GZB4ZZZ'

Warning: Expecting logical in AR5554 / R5554C44: got 'GZB4ZZZ'

Warning: Expecting logical in AS5554 / R5554C45: got 'GZB4ZZZ'

Warning: Expecting logical in AT5554 / R5554C46: got 'GZB4ZZZ'

Warning: Expecting logical in AL5875 / R5875C38: got 'GZB4ZZZ'

Warning: Expecting logical in AL6786 / R6786C38: got 'GZB4ZZZ'

Warning: Expecting logical in AM6786 / R6786C39: got 'GZB4ZZZ'

Warning: Expecting logical in AN6786 / R6786C40: got 'GZB4ZZZ'

Warning: Expecting logical in AO6786 / R6786C41: got 'GZB4ZZZ'

Warning: Expecting logical in AP6786 / R6786C42: got 'GZB4ZZZ'

Warning: Expecting logical in AQ6786 / R6786C43: got 'GZB4ZZZ'

Warning: Expecting logical in AR6786 / R6786C44: got 'GZB4ZZZ'

Warning: Expecting logical in AL9025 / R9025C38: got 'GZB0ZZZ'

Warning: Expecting logical in AM9025 / R9025C39: got 'GZB0ZZZ'

Warning: Expecting logical in AN9025 / R9025C40: got 'GZB0ZZZ'

Warning: Expecting logical in AO9025 / R9025C41: got 'GZB0ZZZ'

Warning: Expecting logical in AL9452 / R9452C38: got 'GZB4ZZZ'

Warning: Expecting logical in AM9452 / R9452C39: got 'GZB4ZZZ'

Warning: Expecting logical in AL10005 / R10005C38: got 'GZB4ZZZ'
```

```
Warning: Expecting logical in AL10637 / R10637C38: got 'GZB4ZZZ'

Warning: Expecting logical in AM10637 / R10637C39: got 'GZB4ZZZ'

Warning: Expecting logical in CX10722 / R10722C102: got 'B2111ZZ'

Warning: Expecting logical in AL11057 / R11057C38: got 'GZB4ZZZ'

Warning: Expecting logical in AL12036 / R12036C38: got 'GZB4ZZZ'

Warning: Expecting logical in AL14754 / R14754C38: got 'GZ50ZZZ'

Warning: Expecting logical in CX14908 / R14908C102: got '0FC44ZZ'
```

```
num_filas <- nrow(SaludMental)
num_columnas <- ncol(SaludMental)

cat("Dimensiones: ", num_filas, " filas x ", num_columnas, " columnas\n", sep = "")
```

```
Dimensiones: 21210 filas x 111 columnas
```

```
head(SaludMental)
```

```
# A tibble: 6 x 111
  `Comunidad Autónoma` Nombre      `Fecha de nacimiento`  Sexo `CCAA Residencia`
  <chr>                <chr>       <dttm>                <dbl> <lgl>
1 ANDALUCÍA            MONICA TIN~ 1951-08-17 00:00:00       2 NA
2 ANDALUCÍA            IRENE RODR~ 1929-03-20 00:00:00       2 NA
3 ANDALUCÍA            JOSE MORIL~ 1976-11-25 00:00:00       1 NA
4 ANDALUCÍA            ELIZABETH ~ 1976-11-10 00:00:00       2 NA
5 ANDALUCÍA            MARIA ENCA~ 1977-04-28 00:00:00       2 NA
6 ANDALUCÍA            ANTONIO BA~ 1986-01-19 00:00:00       1 NA
# i 106 more variables: `Fecha de Ingreso` <dttm>,
#   `Circunstancia de Contacto` <dbl>, `Fecha de Fin Contacto` <chr>,
#   `Tipo Alta` <dbl>, `Estancia Días` <dbl>, `Diagnóstico Principal` <chr>,
#   Categoría <chr>, `Diagnóstico 2` <chr>, `Diagnóstico 3` <chr>,
#   `Diagnóstico 4` <chr>, `Diagnóstico 5` <chr>, `Diagnóstico 6` <chr>,
#   `Diagnóstico 7` <chr>, `Diagnóstico 8` <chr>, `Diagnóstico 9` <chr>,
#   `Diagnóstico 10` <chr>, `Diagnóstico 11` <chr>, `Diagnóstico 12` <chr>, ...
```

0.2.1 Comentarios sobre la carga

- El dataset tiene 21210 filas y 111 columnas.
- Durante la importación inicial del conjunto de datos aparecieron advertencias del tipo *Expecting logical … got '0UT94ZZ'*. R infirió de forma incorrecta que algunas columnas eran lógicas y, al encontrar códigos de texto, generó el aviso. Se forzó la importación de estas columnas como texto para preservar todos los valores originales y evitar pérdidas de información.

0.3 Estructura y tipos de datos

```
glimpse(SaludMental)
```

```
Rows: 21,210
Columns: 111
$ `Comunidad Autónoma`        <chr> "ANDALUCÍA", "ANDALUCÍA", "ANDALUCÍA", "AN~
$ Nombre                      <chr> "MONICA TINEO RODRIGUEZ", "IRENE RODRIGUEZ~
$ `Fecha de nacimiento`       <dttm> 1951-08-17, 1929-03-20, 1976-11-25, 1976-~
$ Sexo                        <dbl> 2, 2, 1, 2, 2, 1, 2, 2, 1, 2, 1, 1, 2, 2, ~
$ `CCAA Residencia`           <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Fecha de Ingreso`          <dttm> 2016-01-01, 2016-01-01, 2016-01-01, 2016-~
$ `Circunstancia de Contacto` <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
$ `Fecha de Fin Contacto`     <chr> "08/01/2016", "08/01/2016", "11/01/2016", ~
$ `Tipo Alta`                 <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
$ `Estancia Días`             <dbl> 7, 7, 10, 26, 17, 57, 3, 19, 24, 1, 7, 5, ~
$ `Diagnóstico Principal`     <chr> "F25.0", "F41.9", "F60.2", "F20.0", "F60.1~
$ Categoría                   <chr> "Esquizofrenia, trastornos esquizotípicos ~
$ `Diagnóstico 2`             <chr> "Z63.79", "I11.9", "F19.288", "C07", "Z88.~
$ `Diagnóstico 3`             <chr> "Z91.19", "I35.8", NA, "F17.210", NA, NA, ~
$ `Diagnóstico 4`             <chr> NA, "E11.9", NA, "F12.20", NA, NA, "E87.6"~
$ `Diagnóstico 5`             <chr> NA, "I87.2", NA, "F14.10", NA, NA, NA, "F1~
$ `Diagnóstico 6`             <chr> NA, "Z95.0", NA, "F10.10", NA, NA, NA, "R4~
$ `Diagnóstico 7`             <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "Z~
$ `Diagnóstico 8`             <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Diagnóstico 9`             <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Diagnóstico 10`            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Diagnóstico 11`            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Diagnóstico 12`            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Diagnóstico 13`            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Diagnóstico 14`            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
```

```
$ `Fecha de Intervención`     <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Procedimiento 1`           <chr> NA, "4B02XSZ", NA, NA, NA, NA, NA, NA, NA,~
$ `Procedimiento 2`           <chr> NA, "B246ZZZ", NA, NA, NA, NA, NA, NA, NA,~
$ `Procedimiento 3`           <chr> NA, "4A02X4Z", NA, NA, NA, NA, NA, NA, NA,~
$ `Procedimiento 4`           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Procedimiento 5`           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Procedimiento 6`           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Procedimiento 7`           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Procedimiento 8`           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Procedimiento 9`           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Procedimiento 10`          <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Procedimiento 11`          <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Procedimiento 12`          <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Procedimiento 13`          <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Procedimiento 14`          <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Procedimiento 15`          <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Procedimiento 16`          <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Procedimiento 17`          <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Procedimiento 18`          <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Procedimiento 19`          <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Procedimiento 20`          <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `GDR AP`                    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `CDM AP`                    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Tipo GDR AP`               <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Valor Peso Español`        <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `GRD APR`                   <dbl> 750, 756, 752, 750, 752, 750, 775, 754, 75~
$ `CDM APR`                   <dbl> 19, 19, 19, 19, 19, 19, 20, 19, 19, 19, 19~
$ `Tipo GDR APR`              <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Valor Peso Americano APR`  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Nivel Severidad APR`       <dbl> 2, 1, 2, 1, 1, 1, 2, 2, 1, 1, 2, 2, 1, 1, ~
$ `Riesgo Mortalidad APR`     <dbl> 1, 2, 1, 2, 1, 1, 1, 2, 1, 1, 2, 1, 1, 1, ~
$ Servicio                    <chr> "PSQ", "CAR", "PSQ", "PSQ", "PSQ", "PSQ", ~
$ Edad                        <dbl> 64, 86, 39, 39, 38, 29, 20, 51, 49, 28, 30~
$ Reingreso                   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Coste APR`                 <dbl> 6340, 2771, 4009, 6073, 3867, 6073, 3996, ~
$ `GDR IR`                    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Tipo GDR IR`               <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Tipo PROCESO IR`           <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ CIE                         <dbl> 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10~
$ `Número de registro anual`  <chr> "8537155.0", "8992115.0", "8998349.0", "88~
$ `Centro Recodificado`       <chr> "-2088791444897189888", "-1166333372325380~
$ `CIP SNS Recodificado`      <chr> "109457269-593755146", "-15897501687813800~
$ `País Nacimiento`           <chr> "724", "ZZZ", "724", "724", "724", "724", ~
```

```
$ `País Residencia`           <chr> "724", "724", "724", "724", "724", "724", ~
$ `Fecha de Inicio contacto`  <chr> "01012016 1622", "01012016 0453", "0101201~
$ `Régimen Financiación`      <chr> "1.0", "1.0", "1.0", "1.0", "1.0", "1.0", ~
$ Procedencia                 <chr> "21.0", "21.0", "21.0", "21.0", "21.0", "2~
$ `Continuidad Asistencial`   <chr> "9.0", "9.0", "9.0", "9.0", "9.0", "9.0", ~
$ `Ingreso en UCI`            <chr> "2.0", "2.0", "2.0", "2.0", "2.0", "2.0", ~
$ `Días UCI`                  <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Diagnóstico 15`            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Diagnóstico 16`            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Diagnóstico 17`            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Diagnóstico 18`            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Diagnóstico 19`            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Diagnóstico 20`            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `POA Diagnóstico Principal` <chr> "S", "S", "S", "S", "S", "S", "S", "S", "S~
$ `POA Diagnóstico 2`         <chr> "E", "S", "S", "S", "E", "S", "S", "S", "S~
$ `POA Diagnóstico 3`         <chr> "S", "S", NA, "S", NA, NA, "S", "S", "S", ~
$ `POA Diagnóstico 4`         <chr> NA, "S", NA, "S", NA, NA, "S", "S", "E", "~
$ `POA Diagnóstico 5`         <chr> NA, "S", NA, "S", NA, NA, NA, "S", NA, "S"~
$ `POA Diagnóstico 6`         <chr> NA, "E", NA, "S", NA, NA, NA, "S", NA, NA,~
$ `POA Diagnóstico 7`         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "E~
$ `POA Diagnóstico 8`         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `POA Diagnóstico 9`         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `POA Diagnóstico 10`        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `POA Diagnóstico 11`        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `POA Diagnóstico 12`        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `POA Diagnóstico 13`        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `POA Diagnóstico 14`        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `POA Diagnóstico 15`        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `POA Diagnóstico 16`        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `POA Diagnóstico 17`        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `POA Diagnóstico 18`        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `POA Diagnóstico 19`        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `POA Diagnóstico 20`        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Procedimiento Externo 1`   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Procedimiento Externo 2`   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Procedimiento Externo 3`   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Procedimiento Externo 4`   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Procedimiento Externo 5`   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Procedimiento Externo 6`   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ `Tipo GRD APR`              <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M~
$ `Peso Español APR`          <dbl> 1.393611, 0.609264, 0.881297, 1.335036, 0.~
$ `Edad en Ingreso`           <dbl> 64, 86, 39, 39, 38, 29, 20, 51, 49, 28, 30~
$ `Mes de Ingreso`            <chr> "2016-01", "2016-01", "2016-01", "2016-01"~
```

```r
tipos <- sapply(SaludMental, function(x) class(x)[1])
as.data.frame(table(tipos))
```

```
      tipos Freq
1 character   70
2   logical   26
3   numeric   13
4   POSIXct    2
```

0.3.1 Interpretación

Panorama general Se identificaron 13 variables numéricas, aunque muchas representan códigos categóricos. Se clasifican en tres grupos:

1. Métricas cuantitativas reales - `Edad` y `Edad en Ingreso` muestran distribuciones casi normales y correlación 1.0; una puede eliminarse por redundancia. - `Estancia Días`, `Coste APR` y `Peso Español APR` presentan fuerte asimetría positiva con casos extremos relevantes para gestión clínica.

2. Códigos categóricos codificados como números - `Sexo`, `Tipo Alta`, `Nivel Severidad APR`, `Riesgo Mortalidad APR`, `GRD APR`, `CDM APR` y similares se concentran en pocos valores enteros. - Deben convertirse en factores con etiquetas descriptivas antes de cualquier modelado.

3. Variables constantes o sin valor analítico - `CIE` es constante (valor 10) y no aporta información.

Relaciones clave - `Estancia Días` y `Coste APR` correlacionan moderadamente (0.231*), lo que respalda su relación esperada. - `Nivel Severidad APR` y `Peso Español APR` tienen correlación positiva (0.251*), reflejando mayor peso en casos severos. - La forma en "L" de las dispersiones sugiere evaluar transformaciones logarítmicas para distribuciones sesgadas.

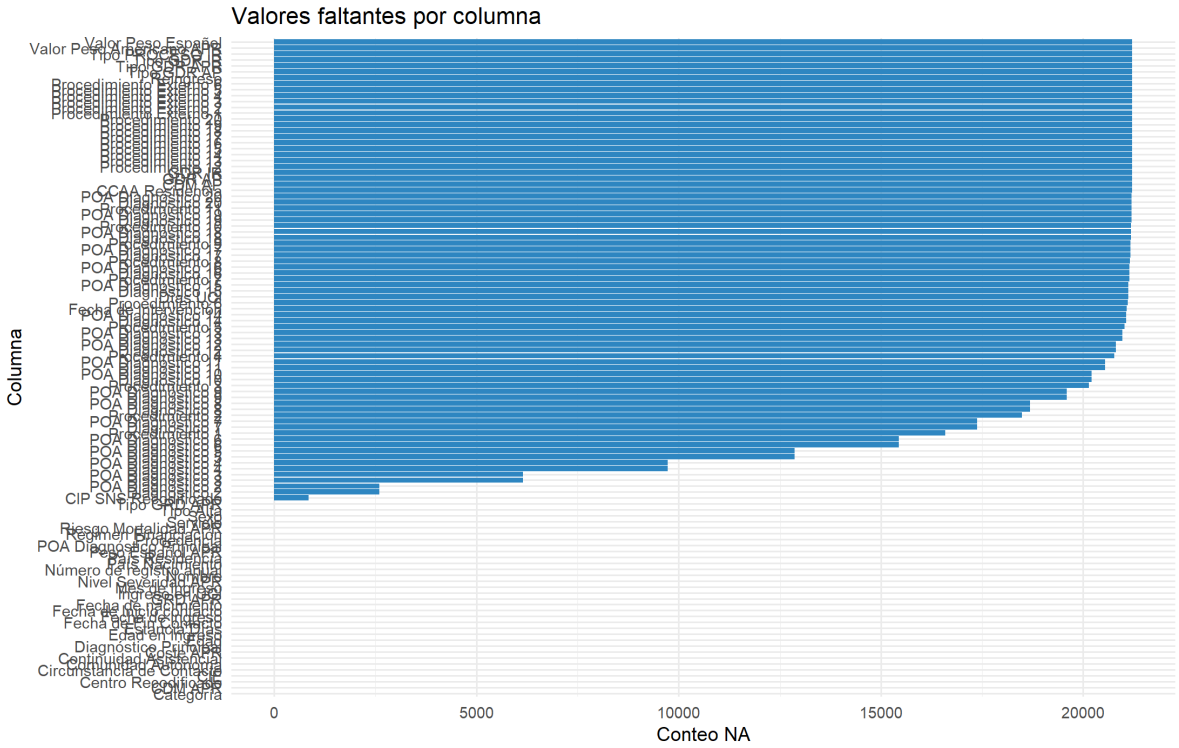
0.4 Valores faltantes (nulos)

```r
na_por_col <- SaludMental %>%
  summarise(across(everything(), ~ sum(is.na(.)))) %>%
  pivot_longer(everything(), names_to = "columna", values_to = "na") %>%
  arrange(desc(na)) %>%
  mutate(pct_na = 100 * na / num_filas)
```

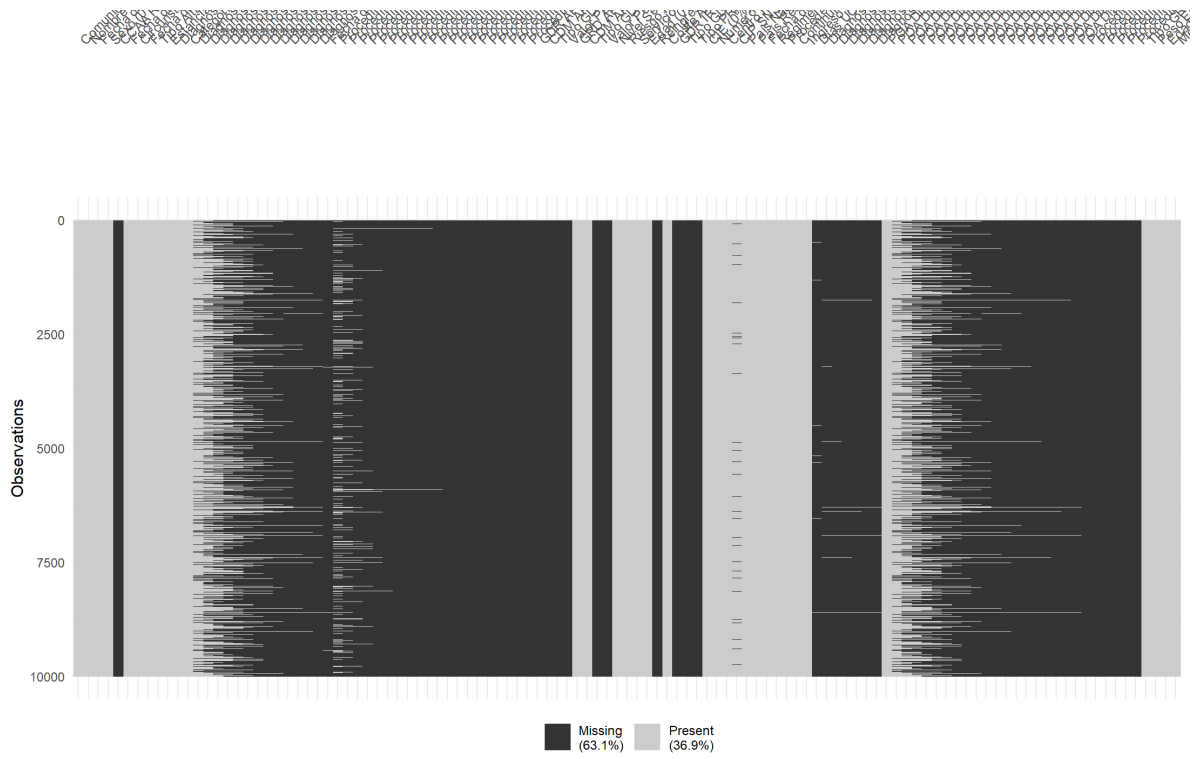
```r
na_top <- head(na_por_col, 3)
na_por_col
```

```
# A tibble: 111 x 3
   columna             na pct_na
   <chr>            <int>  <dbl>
 1 CCAA Residencia  21210    100
 2 Procedimiento 12 21210    100
 3 Procedimiento 13 21210    100
 4 Procedimiento 14 21210    100
 5 Procedimiento 15 21210    100
 6 Procedimiento 16 21210    100
 7 Procedimiento 17 21210    100
 8 Procedimiento 18 21210    100
 9 Procedimiento 19 21210    100
10 Procedimiento 20 21210    100
# i 101 more rows
```

```r
na_plot <- ggplot(na_por_col, aes(x = reorder(columna, na), y = na)) +
  geom_col(fill = "#2E86C1") +
  coord_flip() +
  labs(x = "Columna", y = "Conteo NA", title = "Valores faltantes por columna") +
  theme_minimal(base_size = 13)

na_plot
```




```
sample_size <- min(10000, nrow(SaludMental))
muestra_miss <- dplyr::slice_sample(SaludMental, n = sample_size)
vis_miss(muestra_miss, warn_large_data = FALSE)
```



0.4.1 Interpretación

El análisis revela tres grupos claros de variables según la cantidad de datos faltantes:

1. Columnas completamente vacías (100% NA) - Ejemplos: `CCAA Residencia`, `Procedimiento 12-Procedimiento 20`, `GDR IR`. - Situación: cada variable acumula 21 210 valores faltantes, por lo que no aporta información. - Acción: eliminar estas columnas para reducir ruido y simplificar el modelo de datos.

2. Columnas con alta tasa de valores faltantes (>90% NA) - Ejemplos: `Diagnóstico 3-Diagnóstico 20` y procedimientos secundarios. - Lectura: la mayoría de pacientes no alcanza tantos registros; los NA reflejan esa estructura. - Acción: excluirlas de análisis generales y derivar una métrica del número de diagnósticos/procedimientos por paciente.

Causa Probable: Estas columnas podrían ser placeholders en la base de datos original, campos que nunca se utilizan, o datos que no se extrajeron correctamente durante el proceso de recopilación.

Acción Inmediata: Al no contener ninguna información, estas columnas deben ser eliminadas del dataset. No aportan valor analítico y solo añaden ruido y complejidad.

2. Columnas con Alta Tasa de Valores Faltantes (>90% NA) Existe un segundo grupo de variables, principalmente las relacionadas con diagnósticos y procedimientos secundarios (ej. Diagnóstico 20 hasta Diagnóstico 3), que tienen una tasa de valores faltantes muy alta, aunque no del 100%.

Evidencia: El gráfico vis_miss muestra un patrón claro: la ausencia de datos en Diagnóstico 5, por ejemplo, casi siempre implica la ausencia en Diagnóstico 6 y posteriores.

Interpretación: Esto no es un error, sino una ausencia de datos estructural. Simplemente significa que la mayoría de los pacientes no tienen tantos diagnósticos o procedimientos registrados. Un paciente con solo dos diagnósticos tendrá NA en las columnas Diagnóstico 3 en adelante.

Acción Recomendada: Para la mayoría de los análisis generales, estas columnas con más del 95% de datos faltantes también pueden ser eliminadas. Sin embargo, de este patrón se puede extraer una nueva variable muy útil: un conteo del número total de diagnósticos o procedimientos por paciente, lo que convierte el problema de los datos faltantes en una nueva característica (feature engineering).

3. Variables Clave con Datos Completos o Casi Completos Afortunadamente, las variables que parecen ser centrales para el análisis están en su mayoría completas.

Evidencia: La parte inferior del gráfico de barras muestra que columnas como Categoría, Comunidad Autónoma, Diagnóstico Principal, Sexo, Edad en Ingreso y Estancia Días tienen muy pocos o ningún valor faltante.

Interpretación: Esta es una excelente noticia. Significa que el núcleo del dataset es robusto y fiable, permitiendo realizar análisis descriptivos y de modelado sin necesidad de técnicas complejas de imputación de datos.

0.5 Duplicados

```
num_duplicadas <- sum(duplicated(SaludMental))
pct_duplicadas <- round(100 * num_duplicadas / num_filas, 2)
cat("Filas duplicadas: ", num_duplicadas, " (", pct_duplicadas, "%)\n", sep = "")
```

```
Filas duplicadas: 0 (0%)
```

```
SaludMental %>% filter(duplicated(.)) %>% head(10)
```

```
# A tibble: 0 x 111
# i 111 variables: Comunidad Autónoma <chr>, Nombre <chr>,
#   Fecha de nacimiento <dttm>, Sexo <dbl>, CCAA Residencia <lgl>,
#   Fecha de Ingreso <dttm>, Circunstancia de Contacto <dbl>,
#   Fecha de Fin Contacto <chr>, Tipo Alta <dbl>, Estancia Días <dbl>,
#   Diagnóstico Principal <chr>, Categoría <chr>, Diagnóstico 2 <chr>,
#   Diagnóstico 3 <chr>, Diagnóstico 4 <chr>, Diagnóstico 5 <chr>,
#   Diagnóstico 6 <chr>, Diagnóstico 7 <chr>, Diagnóstico 8 <chr>, ...
```

0.5.1 Interpretación

- No se identificaron filas duplicadas; el porcentaje es 0%.
- Cada registro representa un episodio único, lo que confirma la calidad de la extracción original.
- No se requieren pasos adicionales de limpieza relacionados con duplicados.

0.6 Variables numéricas: resúmenes y gráficos

```
numericas <- SaludMental %>% select(where(is.numeric))
n_num <- ncol(numericas)
cat("Variables numéricas: ", n_num, "\n", sep = "")
```

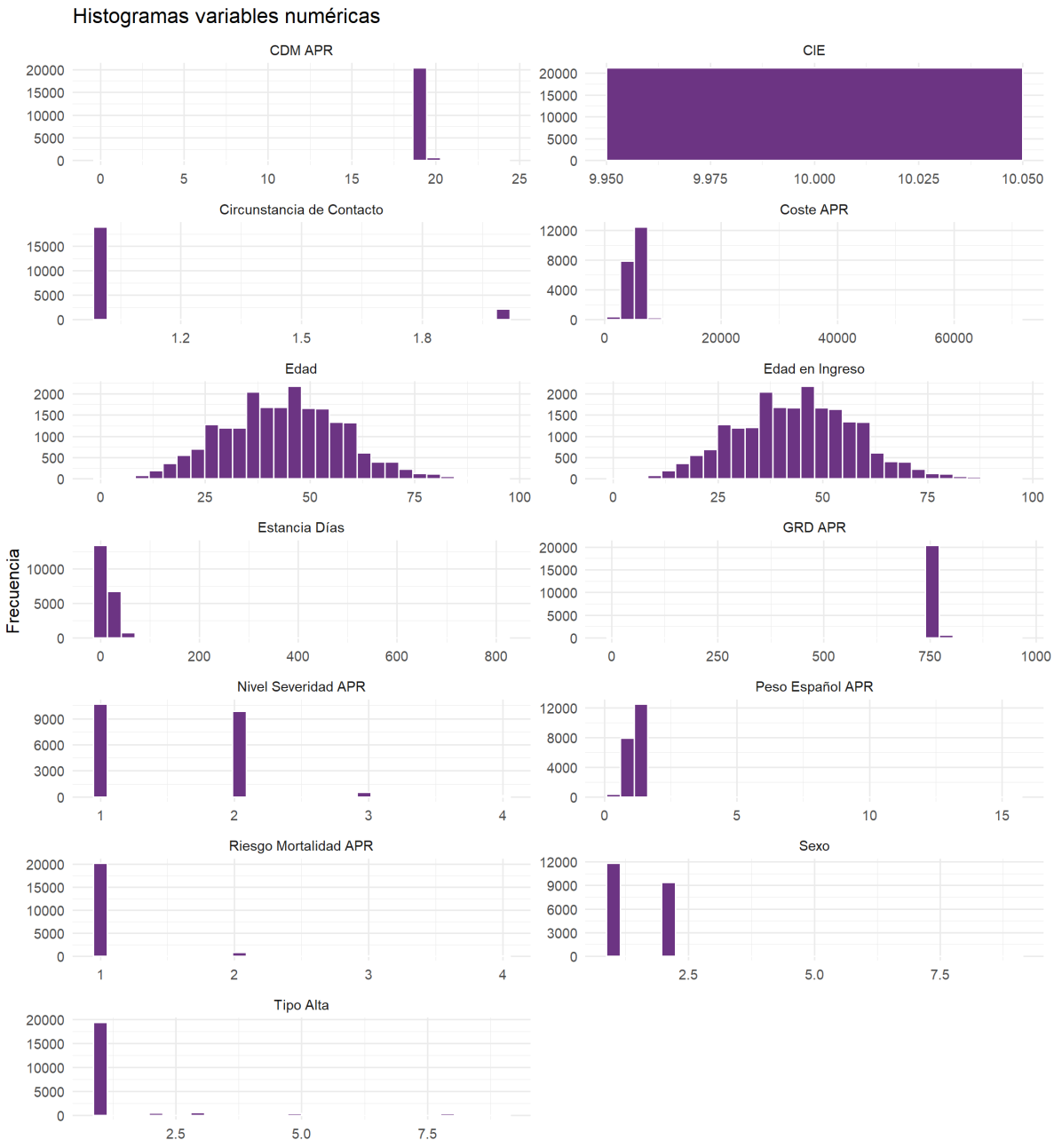
```
Variables numéricas: 13
```

```
summary(numericas)
```

```
      Sexo       Circunstancia de Contacto   Tipo Alta     Estancia Días
 Min.   :1.000   Min.   :1.000             Min.   :1.000   Min.   :  0.00
 1st Qu.:1.000   1st Qu.:1.000             1st Qu.:1.000   1st Qu.:  5.00
 Median :1.000   Median :1.000             Median :1.000   Median : 11.00
 Mean   :1.451   Mean   :1.105             Mean   :1.263   Mean   : 15.46
 3rd Qu.:2.000   3rd Qu.:1.000             3rd Qu.:1.000   3rd Qu.: 19.00
 Max.   :9.000   Max.   :2.000             Max.   :9.000   Max.   :814.00
    GRD APR         CDM APR      Nivel Severidad APR Riesgo Mortalidad APR
 Min.   :  4.0   Min.   : 0.00   Min.   :1.000       Min.   :1.000
 1st Qu.:750.0   1st Qu.:19.00   1st Qu.:1.000       1st Qu.:1.000
 Median :752.0   Median :19.00   Median :1.000       Median :1.000
 Mean   :751.3   Mean   :18.99   Mean   :1.536       Mean   :1.057
 3rd Qu.:753.0   3rd Qu.:19.00   3rd Qu.:2.000       3rd Qu.:1.000
```

```
 Max.   :952.0   Max.   :24.00   Max.   :4.000       Max.   :4.000
     Edad          Coste APR          CIE     Peso Español APR  Edad en Ingreso
 Min.   : 0.00   Min.   : 1496   Min.   :10   Min.   : 0.3298   Min.   : 0.00
 1st Qu.:34.00   1st Qu.: 4228   1st Qu.:10   1st Qu.: 0.9255   1st Qu.:34.00
 Median :44.00   Median : 5988   Median :10   Median : 1.3163   Median :44.00
 Mean   :43.64   Mean   : 5453   Mean   :10   Mean   : 1.1968   Mean   :43.68
 3rd Qu.:53.00   3rd Qu.: 6319   3rd Qu.:10   3rd Qu.: 1.3930   3rd Qu.:53.00
 Max.   :96.00   Max.   :70601   Max.   :10   Max.   :15.5179   Max.   :96.00
```

```r
numericas %>%
  pivot_longer(everything(), names_to = "variable", values_to = "valor") %>%
  ggplot(aes(x = valor)) +
  geom_histogram(bins = 30, fill = "#6C3483", color = "white") +
  facet_wrap(~ variable, scales = "free", ncol = 2) +
  labs(title = "Histogramas variables numéricas", x = NULL, y = "Frecuencia") +
  theme_minimal(base_size = 13)
```



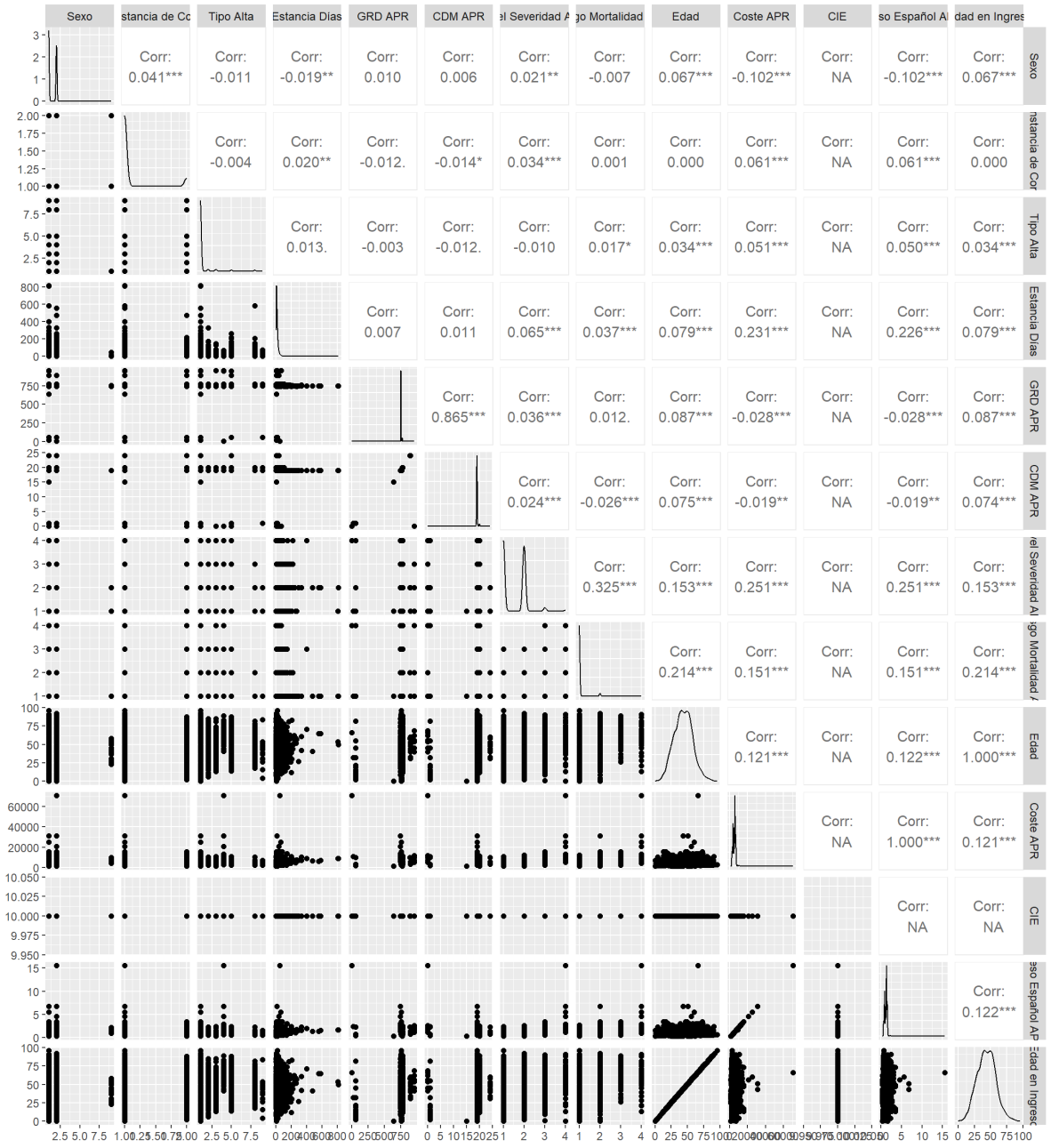

```
numericas %>%
  pivot_longer(everything(), names_to = "variable", values_to = "valor") %>%
  ggplot(aes(x = variable, y = valor)) +
  geom_boxplot(fill = "#F39C12") +
  coord_flip() +
  labs(title = "Boxplots variables numéricas", x = "Variable", y = NULL) +
```

```
  theme_minimal(base_size = 11)
```


```
GGally::ggpairs(
  numericas,
  progress = FALSE,
```

```
  columnLabels = colnames(numericas)
)
```

```
Warning in cor(x, y): La desviación estándar es cero
Warning in cor(x, y): La desviación estándar es cero
Warning in cor(x, y): La desviación estándar es cero
Warning in cor(x, y): La desviación estándar es cero
Warning in cor(x, y): La desviación estándar es cero
Warning in cor(x, y): La desviación estándar es cero
Warning in cor(x, y): La desviación estándar es cero
Warning in cor(x, y): La desviación estándar es cero
Warning in cor(x, y): La desviación estándar es cero
Warning in cor(x, y): La desviación estándar es cero
Warning in cor(x, y): La desviación estándar es cero
Warning in cor(x, y): La desviación estándar es cero
```



Análisis de Variables Numéricas Se identificaron un total de 13 variables con tipo de dato numérico. Sin embargo, un análisis de su distribución y valores revela que no todas son cuantitativas en la práctica. Podemos clasificarlas en tres grupos distintos.

1. Variables Cuantitativas Reales Estas variables representan mediciones continuas y son clave para el análisis.

Edad y Edad en Ingreso: Ambas muestran una distribución casi normal, ligeramente sesgada a la derecha, con un pico en torno a los 40-50 años. El summary y el gráfico de pares (ggpairs) confirman que son prácticamente idénticas (correlación de 1.000), lo que indica que una de ellas es redundante y podría eliminarse.

Estancia Días, Coste APR y Peso Español APR: Estos son los ejemplos más claros de variables cuantitativas. Los histogramas muestran que las tres tienen una fuerte asimetría positiva (sesgo a la derecha). Esto es típico en datos de salud: la mayoría de las estancias y costes son bajos, pero hay unos pocos casos extremos (outliers) con valores muy altos. El boxplot de Coste APR visualiza perfectamente esta característica, con una caja compacta y muchos puntos atípicos.

2. Variables Categóricas Codificadas como Números La mayoría de las variables numéricas en este dataset no son realmente cuantitativas, sino códigos que representan categorías.

Evidencia: Los histogramas para Sexo, Tipo Alta, Nivel Severidad APR, Riesgo Mortalidad APR, GRD APR y CDM APR no muestran una distribución continua, sino barras discretas en valores enteros específicos. Por ejemplo, Sexo se concentra en los valores 1 y 2 (y un valor anómalo en 9), y Nivel Severidad APR en 1, 2, 3 y 4.

Interpretación: Tratar estas variables como números en un modelo (por ejemplo, calculando su media) sería incorrecto. Deben ser convertidas a factores para un análisis adecuado, asignando etiquetas a cada nivel (ej. Sexo: 1 = "Hombre", 2 = "Mujer").

3. Variables Constantes o Inválidas CIE: El summary y el histograma muestran que esta variable tiene un valor constante de 10 para todos los registros. Al no tener variabilidad, no aporta información analítica y no se puede calcular su correlación con otras variables (por eso aparece como NA en el gráfico ggpairs). Esta columna debería ser eliminada.

Correlaciones y Conclusiones Clave El gráfico de pares (ggpairs) nos permite visualizar las relaciones entre las variables:

Redundancia: Como se mencionó, la correlación perfecta (1.000) entre Edad y Edad en Ingreso confirma que son la misma información.

Relaciones Esperadas: Se observa una correlación positiva, aunque moderada, entre Estancia Días y Coste APR (Corr: 0.231*). Lógicamente, estancias más largas tienden a ser más costosas. De manera similar, Nivel Severidad APR tiene una correlación positiva con Peso Español APR (Corr: 0.251*), indicando que casos más severos están asociados a un mayor peso (y por ende, coste).

Necesidad de Transformación: La forma de "L" en los gráficos de dispersión que involucran a Estancia Días y Coste APR es un síntoma claro de su sesgo. Para algunos modelos predictivos, aplicar una transformación logarítmica a estas variables podría ser beneficioso para normalizar su distribución.

0.7 Variables categóricas: frecuencias y gráficos

```
categoricas <- SaludMental %>% select(where(~ is.character(.x) || is.factor(.x)))
n_cat <- ncol(categoricas)
cat("Variables categóricas: ", n_cat, "\n", sep = "")
```

```
Variables categóricas: 70
```

```
vars_interes <- c(
  "Comunidad Autónoma",
  "Nombre",
  "Fecha de Fin Contacto",
  "Diagnóstico Principal",
  "Procedimiento 1",
  "Número de registro anual",
  "Centro Recodificado",
  "País Nacimiento",
  "Mes de Ingreso"
)

categoricas_sel <- categoricas %>% select(all_of(vars_interes))
vars_to_show <- colnames(categoricas_sel)
top_k <- 10

cat("Se grafican ", length(vars_to_show), " variables seleccionadas manualmente.\n", sep = "
```

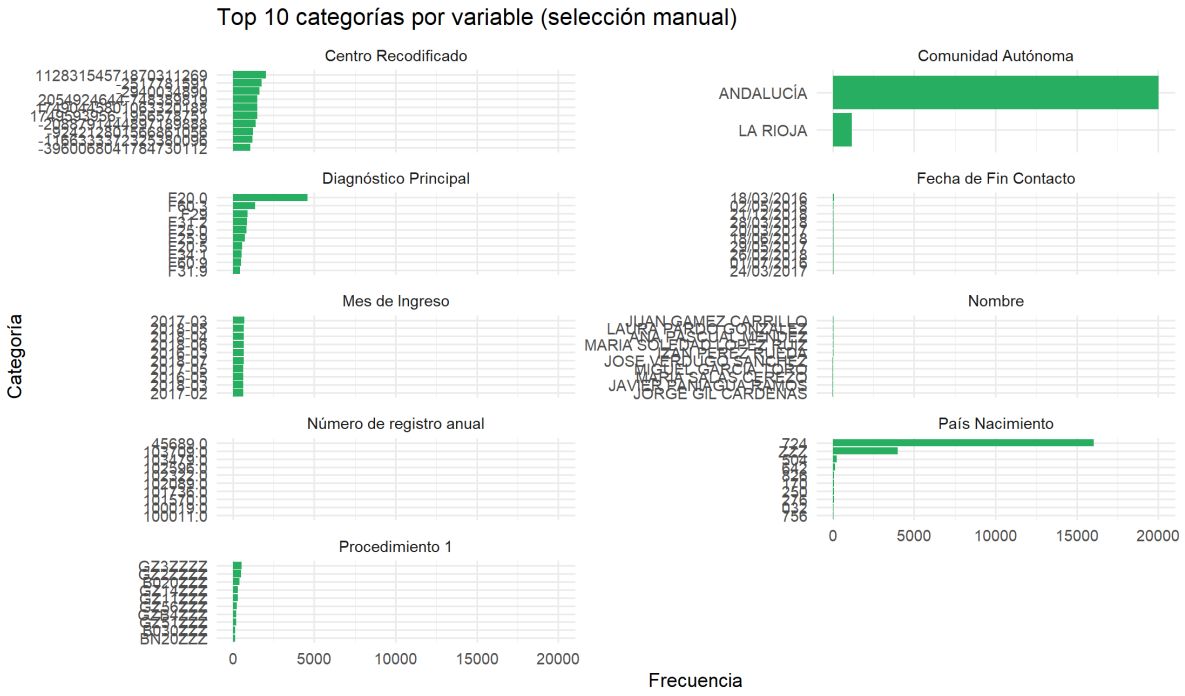
```
Se grafican 9 variables seleccionadas manualmente.
```

```
cat("Variables: ", paste(vars_to_show, collapse = ", "), "\n", sep = "")
```

```
Variables: Comunidad Autónoma, Nombre, Fecha de Fin Contacto, Diagnóstico Principal, Procedi
```

```
if (length(vars_to_show) > 0) {
  counts_cat <- categoricas_sel %>%
    pivot_longer(everything(), names_to = "variable", values_to = "valor") %>%
    filter(!is.na(valor)) %>%
    group_by(variable, valor) %>%
    summarise(n = n(), .groups = "drop")
```

```r
  cat_top <- counts_cat %>%
    group_by(variable) %>%
    slice_max(n, n = top_k, with_ties = FALSE) %>%
    ungroup() %>%
    group_by(variable) %>%
    mutate(valor = forcats::fct_reorder(as.factor(valor), n)) %>%
    ungroup()

  ggplot(cat_top, aes(x = valor, y = n)) +
    geom_col(fill = "#27AE60") +
    coord_flip() +
    facet_wrap(~ variable, scales = "free_y", ncol = 2) +
    labs(
      x = "Categoría",
      y = "Frecuencia",
      title = paste0("Top ", top_k, " categorías por variable (selección manual)")
    ) +
    theme_minimal(base_size = 13)
} else {
  cat("Ninguna de las variables seleccionadas está presente en el dataset.\n")
}
```



0.7.1 Interpretación

El análisis de las nueve variables categóricas seleccionadas revela cuatro hallazgos principales:

1. Sesgo geográfico y demográfico - `Comunidad Autónoma` está dominada por Andalucía; las conclusiones no se generalizan al resto de España. - `País Nacimiento` se concentra en el código `724` (España) y un valor residual `ZZZ`.

2. Perfil clínico dominante - `Diagnóstico Principal` se enfoca en códigos CIE-10 del capítulo F, destacando `F20.9`, `F20.0` y `F31.9`. - `Procedimiento 1` privilegia códigos como `GZZZZZZ` (sin procedimiento) y otros específicos (`B02ZZZZ`, `GZ14ZZZ`).

3. Identificadores de alta cardinalidad - `Nombre`, `Número de registro anual` y `Centro Recodificado` poseen muchos valores únicos; son útiles como identificadores, no para agregaciones.

4. Variables temporales mal tipificadas - `Mes de Ingreso` y `Fecha de Fin Contacto` se almacenan como texto; deben transformarse a formatos de fecha para análisis de tendencias. ## Outliers (regla IQR)

```
outlier_stats <- numericas %>%
  pivot_longer(everything(), names_to = "variable", values_to = "valor") %>%
  group_by(variable) %>%
  summarise(
    q1 = quantile(valor, 0.25, na.rm = TRUE),
    q3 = quantile(valor, 0.75, na.rm = TRUE),
    iqr = q3 - q1,
    lower = q1 - 1.5 * iqr,
    upper = q3 + 1.5 * iqr,
    n = sum(!is.na(valor)),
    outliers = sum(valor < lower | valor > upper, na.rm = TRUE),
    pct_out = round(100 * outliers / pmax(n, 1), 2)
  ) %>% arrange(desc(pct_out))

outlier_stats
```

```
# A tibble: 13 x 9
   variable             q1     q3     iqr    lower  upper     n outliers pct_out
   <chr>             <dbl>  <dbl>   <dbl>    <dbl>  <dbl> <int>    <int>   <dbl>
 1 Circunstancia ~ 1   e+0 1   e0 0          1      1   e0 21210     2221   10.5
 2 Tipo Alta       1   e+0 1   e0 0          1      1   e0 21210     1785    8.42
 3 GRD APR         7.5 e+2 7.53e2 3   e+0  746.    7.58e2 21210     1762    8.31
 4 Estancia Días   5   e+0 1.9 e1 1.4 e+1  -16      4   e1 21210     1229    5.79
 5 Riesgo Mortali~ 1   e+0 1   e0 0          1      1   e0 21210     1013    4.78
```

```
 6 CDM APR          1.9 e+1 1.9 e1 0          19     1.9 e1 21210      808    3.81
 7 Coste APR        4.23e+3 6.32e3 2.09e+3 1092.    9.46e3 21210      191    0.9
 8 Peso Español A~ 9.26e-1 1.39e0 4.67e-1    0.224 2.09e0 21210      179    0.84
 9 Edad             3.4 e+1 5.3 e1 1.9 e+1    5.5   8.15e1 21210      149    0.7
10 Edad en Ingreso 3.4 e+1 5.3 e1 1.9 e+1    5.5   8.15e1 21210      149    0.7
11 Nivel Severida~ 1   e+0 2   e0 1   e+0   -0.5   3.5 e0 21210      149    0.7
12 Sexo            1   e+0 2   e0 1   e+0   -0.5   3.5 e0 21210       25    0.12
13 CIE             1   e+1 1   e1 0          10     1   e1 21210        0    0
```

```r
top_var <- outlier_stats$variable[1]
top_pct <- outlier_stats$pct_out[1]
cat("Mayor % de outliers en: ", top_var, " (", top_pct, "%)\n", sep = "")
```

```
Mayor % de outliers en: Circunstancia de Contacto (10.47%)
```

0.7.2 Interpretación

El análisis IQR diferencia dos tipos de resultados:

1. Falsos positivos en códigos categóricos - `Circunstancia de Contacto`, `Tipo Alta`, `GRD APR`, `Riesgo Mortalidad APR` y similares aparecen con altos porcentajes de outliers. - La causa es que el IQR se aplica a columnas de códigos discretos con baja variabilidad (por ejemplo, solo valores 1 y 2). - Acción: reclasificar estas variables como factores y excluirlas del análisis de outliers.

2. Outliers reales en métricas cuantitativas - `Estancia Días` (5.8%), `Coste APR` (0.9%) y `Peso Español APR` (0.8%) contienen casos extremos auténticos. - Representan estancias prolongadas y episodios de alto coste; conviene analizarlos detalladamente antes de decidir cualquier tratamiento. - `Edad` y `Edad en Ingreso` muestran outliers en extremos etarios, coherentes con la distribución.

Conclusión Los outliers validan la necesidad de diferenciar variables categóricas y cuantitativas, y subrayan que los casos extremos en costes y estancias son cruciales para la toma de decisiones clínicas y de gestión.

1 Análisis univariable

A continuación, analizaremos 1 a 1 columnas de gran interés.

1.1 Edad: análisis univariado

La edad de los pacientes es un factor estructural para comprender la carga asistencial. A continuación se valida el campo, se depuran valores no plausibles y se resumen sus características centrales.

```
datos_edad <- SaludMental %>%
  mutate(
    Edad_texto = as.character(`Edad`),
    Edad_numerica = suppressWarnings(as.numeric(Edad_texto))
  )

n_total_edad <- nrow(datos_edad)
n_no_numerico_edad <- sum(is.na(datos_edad$Edad_numerica) & !is.na(datos_edad$Edad_texto))
n_negativas_edad <- sum(!is.na(datos_edad$Edad_numerica) & datos_edad$Edad_numerica < 0)
n_mayores_100_edad <- sum(!is.na(datos_edad$Edad_numerica) & datos_edad$Edad_numerica > 100)

resumen_limpieza_edad <- tibble(
  indicador = c(
    "Observaciones totales",
    "Conversiones no numéricas a NA",
    "Edades negativas marcadas como NA",
    "Edades > 100 marcadas como NA"
  ),
  conteo = c(
    n_total_edad,
    n_no_numerico_edad,
    n_negativas_edad,
    n_mayores_100_edad
  )
)

datos_edad_limpios <- datos_edad %>%
  mutate(
    Edad_limpia = dplyr::case_when(
      is.na(Edad_numerica) ~ NA_real_,
      Edad_numerica < 0 ~ NA_real_,
      Edad_numerica > 100 ~ NA_real_,
      TRUE ~ Edad_numerica
    )
  )

edad_metricas <- datos_edad_limpios %>%
```

```
  summarise(
    conteo = sum(!is.na(Edad_limpia)),
    media = mean(Edad_limpia, na.rm = TRUE),
    mediana = median(Edad_limpia, na.rm = TRUE),
    desviacion = sd(Edad_limpia, na.rm = TRUE),
    minimo = min(Edad_limpia, na.rm = TRUE),
    q1 = quantile(Edad_limpia, 0.25, na.rm = TRUE),
    q3 = quantile(Edad_limpia, 0.75, na.rm = TRUE),
    maximo = max(Edad_limpia, na.rm = TRUE)
  ) %>%
  mutate(across(where(is.numeric), ~ round(., 2)))

knitr::kable(resumen_limpieza_edad, caption = "Depuración de la variable Edad")
```

Table 1: Depuración de la variable Edad

indicador	conteo
Observaciones totales	21210
Conversiones no numéricas a NA	0
Edades negativas marcadas como NA	0
Edades > 100 marcadas como NA	0

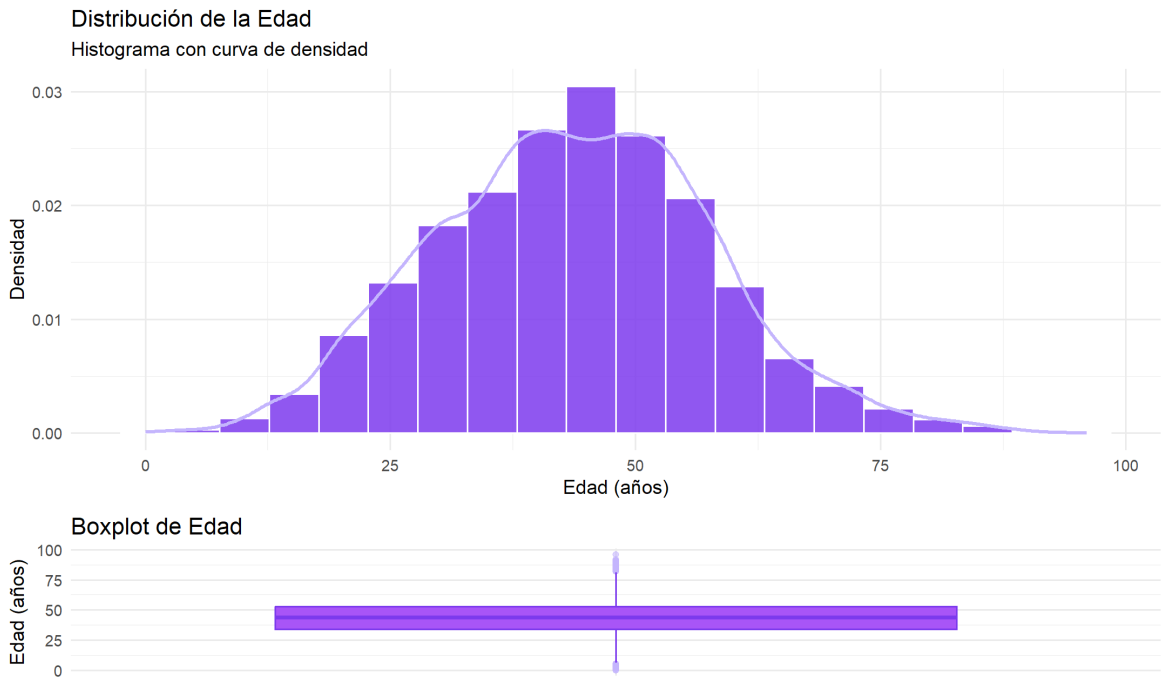
```
knitr::kable(edad_metricas, caption = "Estadísticas descriptivas de Edad (datos validados)")
```

Table 2: Estadísticas descriptivas de Edad (datos validados)

conteo	media	mediana	desviacion	minimo	q1	q3	maximo
21210	43.64	44	14.11	0	34	53	96

```
edad_para_graficos <- datos_edad_limpios %>%
  filter(!is.na(Edad_limpia))

hist_edad <- ggplot(edad_para_graficos, aes(x = Edad_limpia)) +
  geom_histogram(
    aes(y = after_stat(density)),
    bins = 20,
    fill = "#7C3AED",
    color = "white",
    alpha = 0.85
```

```
  ) +
  geom_density(color = "#C4B5FD", linewidth = 1.1) +
  labs(
    title = "Distribución de la Edad",
    subtitle = "Histograma con curva de densidad",
    x = "Edad (años)",
    y = "Densidad"
  ) +
  theme_minimal(base_size = 13)

box_edad <- ggplot(edad_para_graficos, aes(x = "", y = Edad_limpia)) +
  geom_boxplot(
    fill = "#A855F7",
    color = "#7C3AED",
    outlier.color = "#C4B5FD",
    outlier.alpha = 0.6
  ) +
  labs(
    title = "Boxplot de Edad",
    x = NULL,
    y = "Edad (años)"
  ) +
  theme_minimal(base_size = 13) +
  theme(axis.text.x = element_blank(), axis.ticks.x = element_blank())

hist_edad / box_edad + plot_layout(heights = c(3, 1))
```



1.1.1 Interpretación de la Distribución de la Edad

La visualización combinada del histograma con curva de densidad y el diagrama de caja (boxplot) ofrece una visión completa y coherente sobre la distribución de la variable Edad.

- **Distribución Simétrica y Normal:** El histograma muestra una clara forma de campana, confirmada por la curva de densidad superpuesta. Esto indica que la distribución de la edad es aproximadamente normal y simétrica. La mayoría de los casos se agrupan en torno a un valor central, y la frecuencia disminuye de manera uniforme hacia los extremos más jóvenes y más ancianos.
- **Convergencia de las Medidas Centrales:** A diferencia de una distribución asimétrica, en este caso la media, la mediana y la moda son muy similares. El pico del histograma (moda) y la línea central del diagrama de caja (mediana) se sitúan alrededor de los 45-50 años. Esta simetría hace que la media aritmética sea una medida fiable y representativa del "caso típico" en cuanto a la edad.
- **Concentración en la Edad Adulta:** El diagrama de caja revela que el 50% central de los individuos (el rango intercuartílico) se encuentra aproximadamente entre los 38 y los 60 años. Esto indica que el núcleo de la población analizada corresponde a adultos de mediana edad.

- **Presencia de Outliers no Influyentes:** El diagrama de caja señala la existencia de algunos valores atípicos (outliers), representados como puntos individuales por encima y por debajo de los "bigotes". Estos casos representan individuos en los extremos de edad (muy jóvenes o muy mayores) que se desvían de la norma. Sin embargo, su escaso número no altera la forma fundamental de la distribución normal.

En resumen: La edad de la población estudiada sigue un patrón de distribución normal, con una fuerte concentración de individuos en la mediana edad. Esta simetría sugiere que no existen sesgos significativos hacia grupos de edad particulares (ni predominantemente jóvenes ni ancianos). Cualquier análisis o modelo que utilice la edad como variable predictora puede beneficiarse de esta distribución bien comportada, ya que cumple con los supuestos de normalidad que muchos métodos estadísticos requieren.

1.2 Sexo: distribución de episodios

El campo `Sexo` se almacena como código numérico. Se harmonizan las etiquetas, se calculan frecuencias absolutas y relativas, y se visualiza la participación de cada grupo.

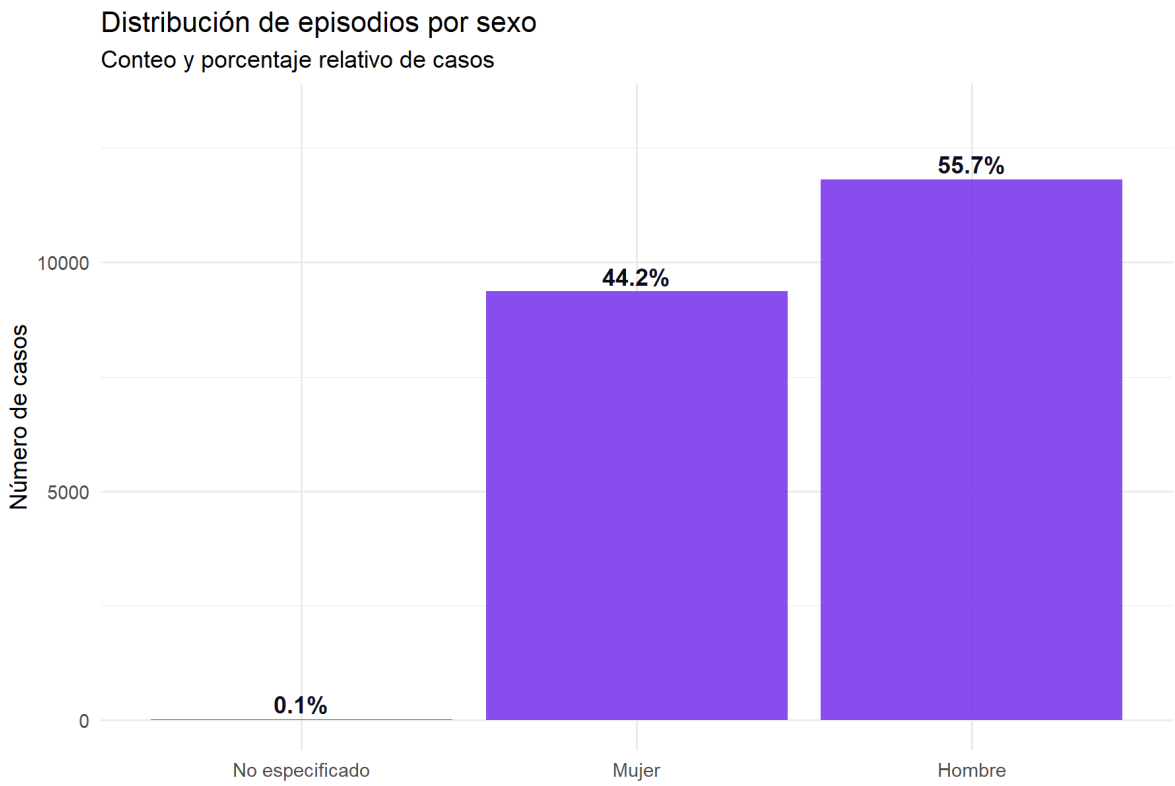
```
tabla_sexo <- SaludMental %>%
  mutate(
    sexo_original = as.character(Sexo),
    sexo_num = suppressWarnings(as.numeric(sexo_original)),
    sexo_etiqueta = case_when(
      is.na(sexo_num) ~ "No disponible",
      sexo_num == 1 ~ "Hombre",
      sexo_num == 2 ~ "Mujer",
      sexo_num == 9 ~ "No especificado",
      TRUE ~ "Valor no mapeado"
    )
  ) %>%
  count(sexo_etiqueta, name = "conteo") %>%
  mutate(
    porcentaje = round(100 * conteo / sum(conteo), 1)
  ) %>%
  arrange(desc(conteo)) %>%
  mutate(
    etiqueta = paste0(conteo, " casos (", porcentaje, "%)"),
    sexo_etiqueta = forcats::fct_reorder(sexo_etiqueta, conteo)
  )

tabla_sexo %>%
  select(Sexo = sexo_etiqueta, Conteo = conteo, Porcentaje = porcentaje) %>%
```

```
  arrange(desc(Conteo)) %>%
  knitr::kable(caption = "Frecuencia absoluta y relativa de Sexo")
```

Table 3: Frecuencia absoluta y relativa de Sexo

Sexo	Conteo	Porcentaje
Hombre	11817	55.7
Mujer	9368	44.2
No especificado	25	0.1

```
ggplot(tabla_sexo, aes(x = sexo_etiqueta, y = conteo)) +
  geom_col(fill = "#7C3AED", alpha = 0.9) +
  geom_text(
    aes(label = paste0(porcentaje, "%")),
    vjust = -0.35,
    color = "#0D0C1D",
    fontface = "bold"
  ) +
  expand_limits(y = max(tabla_sexo$conteo) * 1.12) +
  labs(
    title = "Distribución de episodios por sexo",
    subtitle = "Conteo y porcentaje relativo de casos",
    x = NULL,
    y = "Número de casos"
  ) +
  theme_minimal(base_size = 13)
```



1.2.1 Interpretación de la Distribución por Sexo

El gráfico de barras muestra de forma directa y concisa la distribución de los episodios clínicos según el sexo de los individuos.

- **Clara Disparidad entre Sexos:** La observación más inmediata es la diferencia en la frecuencia de casos entre hombres y mujeres. El gráfico evidencia que no hay una distribución equitativa.
- **Predominio Masculino:** Los hombres representan la mayoría de los casos, constituyendo un **55.7%** del total. En contraste, las mujeres conforman el **44.2%**. Esta diferencia de más de 11 puntos porcentuales indica que, en el contexto de los datos analizados, los hombres experimentan estos episodios con una frecuencia notablemente mayor.
- **Calidad del Dato:** La categoría "No especificado" es prácticamente inexistente, con solo un **0.1%** de los casos. Esto es un fuerte indicador de la alta calidad y completitud de los datos para esta variable, lo que confiere una gran fiabilidad a la comparación entre los dos sexos.

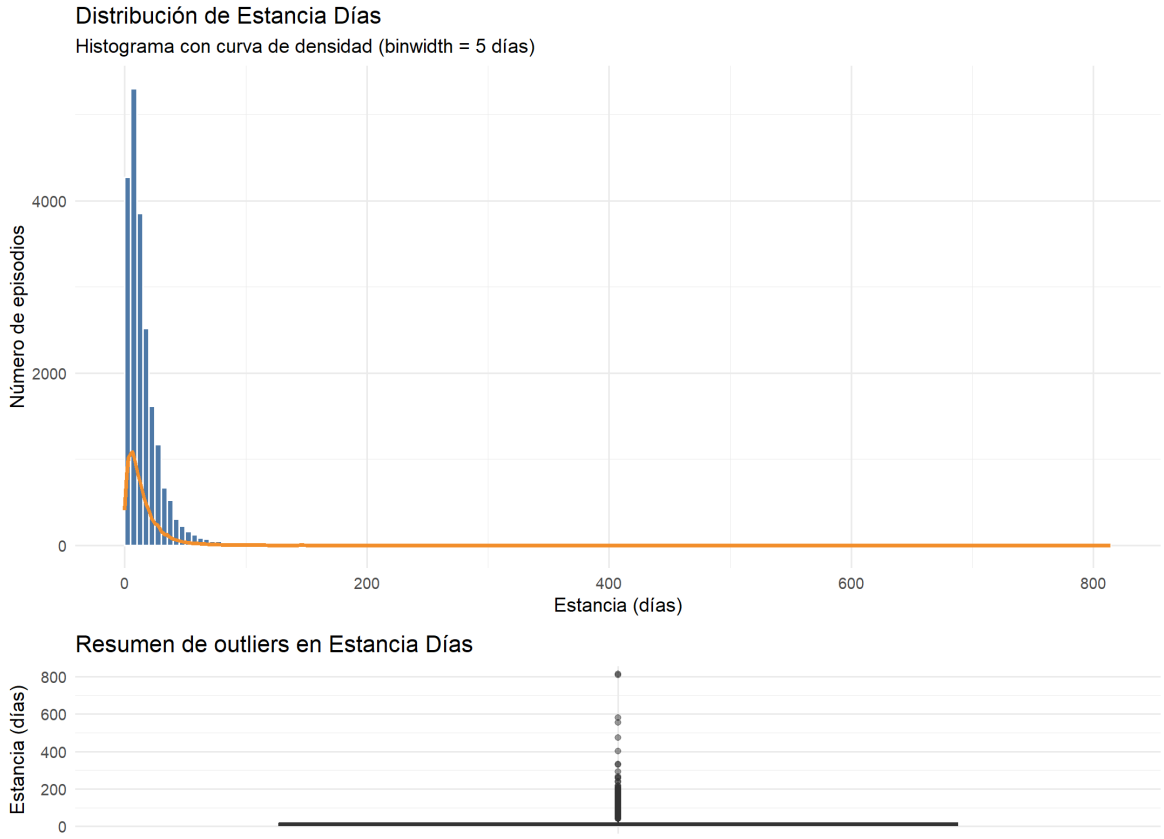
En resumen: El conjunto de datos presenta un claro sesgo de sexo, con una mayor incidencia de episodios en la población masculina. Esta disparidad es un hallazgo fundamental y debe ser considerado como un factor relevante en cualquier análisis posterior. Investigar las posibles causas de esta diferencia podría ser un punto clave para entender la naturaleza del fenómeno estudiado.

1.3 Estancia Días: exploración gráfica combinada

La duración de las estancias hospitalarias es una métrica crítica para detectar posibles ineficiencias o casos clínicos complejos. La combinación de histogramas y diagramas de caja permite identificar simultáneamente la forma de la distribución y los valores atípicos que podrían requerir revisión clínica o protocolos diferenciales.

```
datos_estancia <- SaludMental %>%
  filter(!is.na(`Estancia Días`))

hist_estancia <- ggplot(datos_estancia, aes(x = `Estancia Días`)) +
  geom_histogram(
    binwidth = 5,
    fill = "#4e79a7",
    color = "white",
    boundary = 0,
    closed = "left"
  ) +
  geom_density(aes(y = after_stat(count)), color = "#f28e2b", linewidth = 1.2) +
  labs(
    title = "Distribución de Estancia Días",
    subtitle = "Histograma con curva de densidad (binwidth = 5 días)",
    x = "Estancia (días)",
    y = "Número de episodios"
  ) +
  theme_minimal(base_size = 13)

box_estancia <- ggplot(datos_estancia, aes(x = "Estancia Días", y = `Estancia Días`)) +
  geom_boxplot(fill = "#76b7b2", outlier.alpha = 0.5) +
  labs(
    title = "Resumen de outliers en Estancia Días",
    x = NULL,
    y = "Estancia (días)"
  ) +
  theme_minimal(base_size = 13) +
  theme(axis.text.x = element_blank(), axis.ticks.x = element_blank())
```

```
hist_estancia / box_estancia +
  plot_layout(heights = c(3, 1))
```



1.3.1 Interpretación de la Distribución de la Estancia en Días

La visualización combinada del histograma y el diagrama de caja ofrece una visión clara y contundente sobre la naturaleza de la variable `Estancia Días`.

1. **Fuerte Asimetría Positiva:** El histograma muestra una distribución con un marcado sesgo hacia la izquierda. Esto significa que la gran mayoría de las estancias hospitalarias son de corta duración, con una alta concentración de casos en los primeros días tras el ingreso.
2. **Presencia de Outliers Relevantes:** La larga cola hacia la derecha en el histograma, confirmada por los numerosos puntos individuales en el diagrama de caja, indica la presencia de valores atípicos significativos. Estos puntos representan un número reducido de pacientes cuyas estancias son excepcionalmente largas en comparación con la norma.

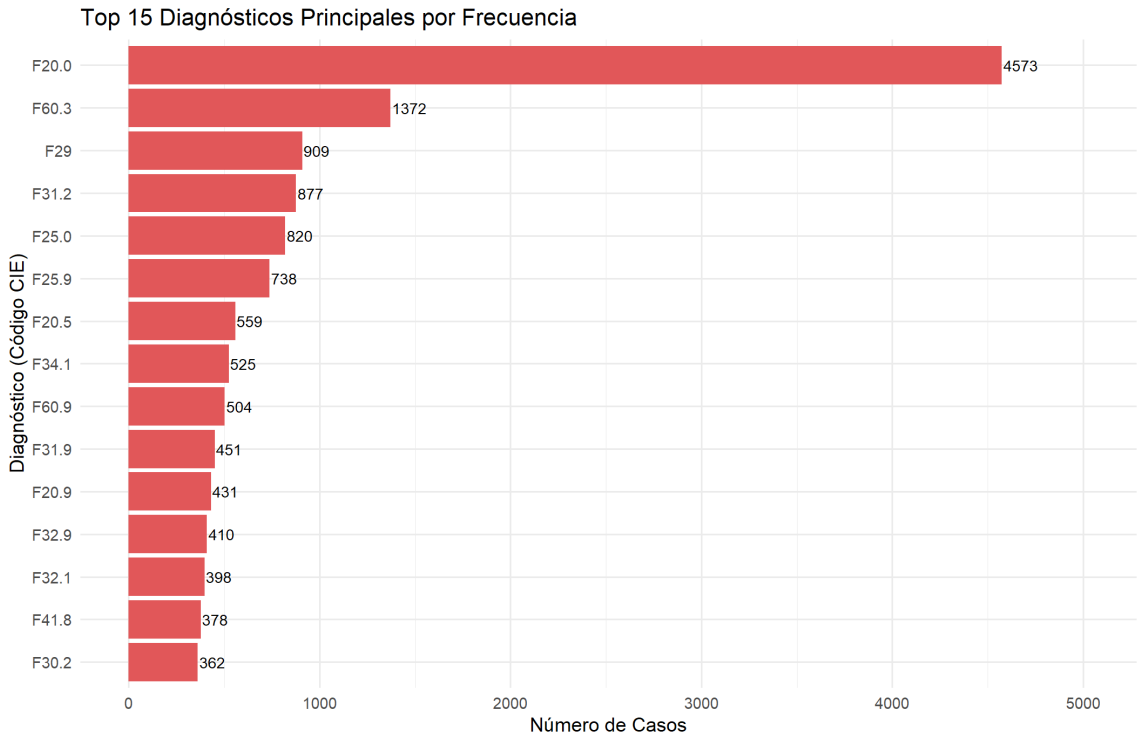
3. **La Media como Medida Engañosa:** Debido a esta asimetría, la media aritmética de los días de estancia se verá fuertemente influenciada por estos valores extremos y no será un buen representante del "caso típico". El diagrama de caja muestra que la mediana (la línea dentro de la caja), que representa el valor central del 50% de los datos, es mucho más baja. Por lo tanto, **la mediana es una métrica más robusta y fiable** para describir la duración de una estancia habitual.

En resumen: La mayoría de los pacientes permanecen hospitalizados por un periodo corto. Sin embargo, existe un subgrupo de pacientes con estancias muy prolongadas que, aunque son una minoría, son cruciales para el análisis, ya que probablemente representen los casos más complejos, costosos o con mayores comorbilidades. Cualquier análisis posterior o modelo predictivo deberá tener en cuenta esta distribución no normal para evitar conclusiones erróneas.

1.4 Diagnóstico Principal: prevalencias clave

El diagnóstico nominal define la naturaleza clínica de cada episodio. Identificar los códigos CIE-10 más frecuentes ayuda a priorizar líneas asistenciales, enfocar estudios específicos y anticipar qué patologías dominarán cualquier modelo predictivo posterior.

```
diagnosticos_top <- SaludMental %>%
  filter(!is.na(`Diagnóstico Principal`)) %>%
  count(`Diagnóstico Principal`, name = "casos") %>%
  arrange(desc(casos)) %>%
  slice_head(n = 15) %>%
  mutate(`Diagnóstico Principal` = forcats::fct_reorder(`Diagnóstico Principal`, casos))

ggplot(diagnosticos_top, aes(x = `Diagnóstico Principal`, y = casos)) +
  geom_col(fill = "#e15759") +
  geom_text(aes(label = casos), hjust = -0.05, size = 3.5) +
  coord_flip(clip = "off") +
  expand_limits(y = max(diagnosticos_top$casos) * 1.1) +
  labs(
    title = "Top 15 Diagnósticos Principales por Frecuencia",
    x = "Diagnóstico (Código CIE)",
    y = "Número de Casos"
  ) +
  theme_minimal(base_size = 13) +
  theme(plot.margin = margin(10, 30, 10, 10))
```



1.4.1 Interpretación de los diagnósticos predominantes

La gráfica de barras horizontales ofrece una visión clara e inequívoca sobre la distribución de los diagnósticos más frecuentes en el conjunto de datos.

- **Dominancia de un Único Diagnóstico:** La visualización revela una concentración extraordinaria en el código **F20.0 (Esquizofrenia paranoide)**, que, con 4,573 casos, supera por más del triple al segundo diagnóstico más común. Esta abrumadora frecuencia lo posiciona como el problema de salud central y más prevalente dentro de esta población de pacientes.
- **Distribución de Pareto:** Se observa una distribución muy asimétrica, similar al principio de Pareto, donde un número muy reducido de diagnósticos representa una proporción masiva de los casos totales. Después del F20.0, la frecuencia desciende bruscamente con el **F60.3 (Trastorno de inestabilidad emocional de la personalidad)** y luego de manera más gradual, formando una "larga cola" de diagnósticos menos comunes.
- **Agrupación por Patologías:** Es notable que la mayoría de los códigos en el top 15 pertenecen a un espectro limitado de trastornos mentales graves. Predominan los códigos del grupo **F20-F29 (Esquizofrenia, trastorno esquizotípico y trastornos de ideas delirantes)** y del grupo **F30-F39 (Trastornos del humor)**. Esto sugiere

que el conjunto de datos probablemente proviene de una unidad o servicio especializado en psicosis y trastornos afectivos graves.

En resumen: El análisis está fuertemente dominado por el diagnóstico de esquizofrenia paranoide. Esto indica que cualquier estrategia de gestión, asignación de recursos o análisis predictivo debe tener como eje central el abordaje de esta patología. La prevalencia de otros trastornos psicóticos y de la personalidad, aunque menor, señala las comorbilidades y diagnósticos secundarios más probables en esta población específica. Ignorar esta marcada asimetría conduciría a conclusiones generalistas y poco representativas de la realidad clínica que reflejan los datos.

#Análisis bivariable

1.5 Edad y Sexo: análisis comparativo

La comparación conjunta de edad y sexo permite detectar diferencias en el perfil etario de los episodios y anticipar posibles segmentaciones clínicas.

```
datos_edad_sexo <- datos_edad_limpios %>%
  mutate(
    sexo_original = as.character(Sexo),
    sexo_num = suppressWarnings(as.numeric(sexo_original)),
    sexo_factor = case_when(
      sexo_num == 1 ~ "Hombre",
      sexo_num == 2 ~ "Mujer",
      sexo_num == 9 ~ "No especificado",
      is.na(sexo_num) ~ "No disponible",
      TRUE ~ "Valor no mapeado"
    )
  ) %>%
  mutate(
    sexo_factor = forcats::fct_relevel(
      factor(sexo_factor),
      "Hombre", "Mujer", "No especificado", "No disponible", "Valor no mapeado"
    )
  )
```

```
Warning: There was 1 warning in `mutate()`.
i In argument: `sexo_factor = forcats::fct_relevel(...)`.
Caused by warning:
! 2 unknown levels in `f`: No disponible and Valor no mapeado
```

```r
datos_edad_sexo_filtrado <- datos_edad_sexo %>%
  filter(!is.na(Edad_limpia), !is.na(sexo_factor), sexo_factor != "Valor no mapeado")

registros_utilizados <- nrow(datos_edad_sexo_filtrado)
registros_excluidos <- nrow(datos_edad_sexo) - registros_utilizados

cat(
  "Registros utilizados en el análisis bivariado: ", registros_utilizados,
  " | Registros excluidos por NA o códigos no mapeados: ", registros_excluidos,
  "\n",
  sep = ""
)
```

```
Registros utilizados en el análisis bivariado: 21210 | Registros excluidos por NA o códigos n
```

```r
estadisticas_edad_sexo <- datos_edad_sexo_filtrado %>%
  group_by(sexo_factor) %>%
  summarise(
    n = n(),
    media = mean(Edad_limpia),
    mediana = median(Edad_limpia),
    desviacion = sd(Edad_limpia),
    minimo = min(Edad_limpia),
    maximo = max(Edad_limpia),
    .groups = "drop"
  ) %>%
  mutate(across(where(is.numeric), ~ round(., 2)))

media_hombre <- estadisticas_edad_sexo %>% filter(sexo_factor == "Hombre") %>% pull(media)
media_mujer <- estadisticas_edad_sexo %>% filter(sexo_factor == "Mujer") %>% pull(media)
mediana_hombre <- estadisticas_edad_sexo %>% filter(sexo_factor == "Hombre") %>% pull(mediana
mediana_mujer <- estadisticas_edad_sexo %>% filter(sexo_factor == "Mujer") %>% pull(mediana)
sd_hombre <- estadisticas_edad_sexo %>% filter(sexo_factor == "Hombre") %>% pull(desviacion)
sd_mujer <- estadisticas_edad_sexo %>% filter(sexo_factor == "Mujer") %>% pull(desviacion)

dif_media <- if (!is.na(media_hombre) && !is.na(media_mujer)) round(abs(media_hombre - media_
dif_mediana <- if (!is.na(mediana_hombre) && !is.na(mediana_mujer)) round(abs(mediana_hombre

cola_maxima <- datos_edad_sexo_filtrado %>%
  filter(sexo_factor %in% c("Hombre", "Mujer")) %>%
  group_by(sexo_factor) %>%
```

```
  summarise(maximo = max(Edad_limpia), .groups = "drop")

knitr::kable(
  estadisticas_edad_sexo,
  col.names = c(
    "Sexo",
    "Casos",
    "Media",
    "Mediana",
    "Desviación estándar",
    "Mínimo",
    "Máximo"
  ),
  caption = "Estadísticas descriptivas de Edad por Sexo"
)
```

Table 4: Estadísticas descriptivas de Edad por Sexo

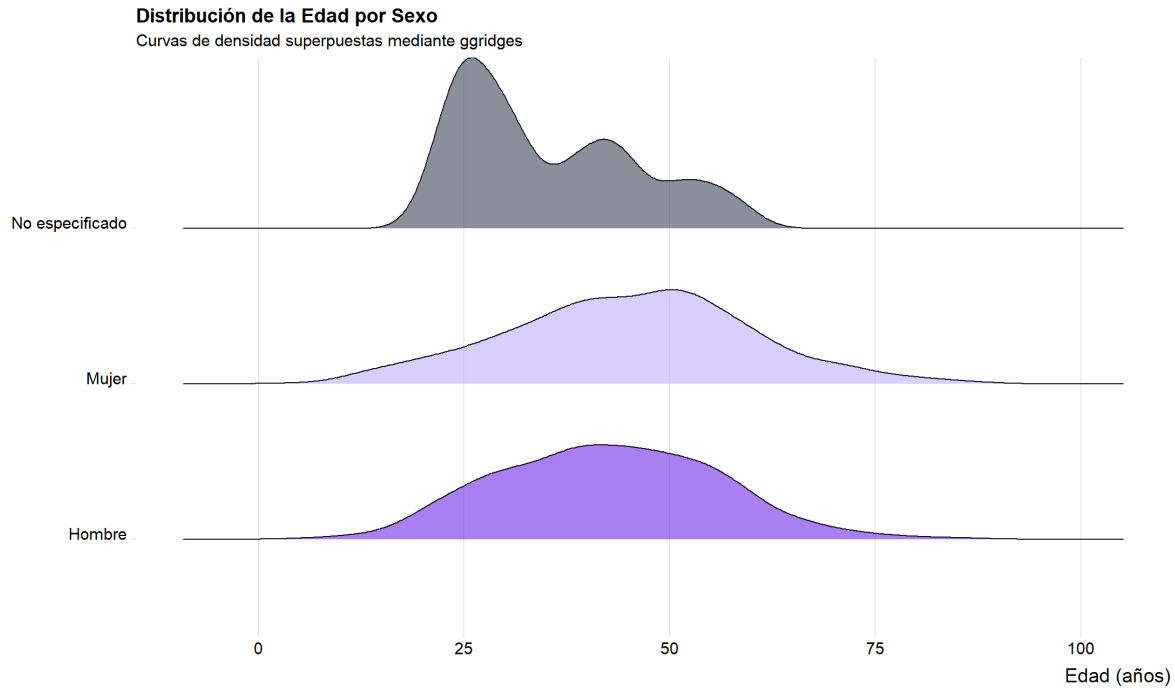
Sexo	Casos	Media	Mediana	Desviación estándar	Mínimo	Máximo
Hombre	11817	42.57	42	13.42	0	96
Mujer	9368	45.02	45	14.82	0	92
No especificado	25	35.16	31	10.89	23	58

```
paleta_sexo <- c(
  "Hombre" = "#7C3AED",
  "Mujer" = "#C4B5FD",
  "No especificado" = "#4B5563",
  "No disponible" = "#9CA3AF"
)

ggplot(datos_edad_sexo_filtrado, aes(x = Edad_limpia, y = sexo_factor, fill = sexo_factor)) +
  geom_density_ridges(
    alpha = 0.65,
    scale = 1.1,
    color = "#0D0C1D",
    size = 0.4
  ) +
  scale_fill_manual(values = paleta_sexo, guide = "none") +
  labs(
    title = "Distribución de la Edad por Sexo",
    subtitle = "Curvas de densidad superpuestas mediante ggridges",
```

```
    x = "Edad (años)",
    y = NULL
  ) +
  theme_ridges(font_size = 13, grid = TRUE) +
  theme(
    panel.grid.major.y = element_blank(),
    axis.title.y = element_blank()
  )
```

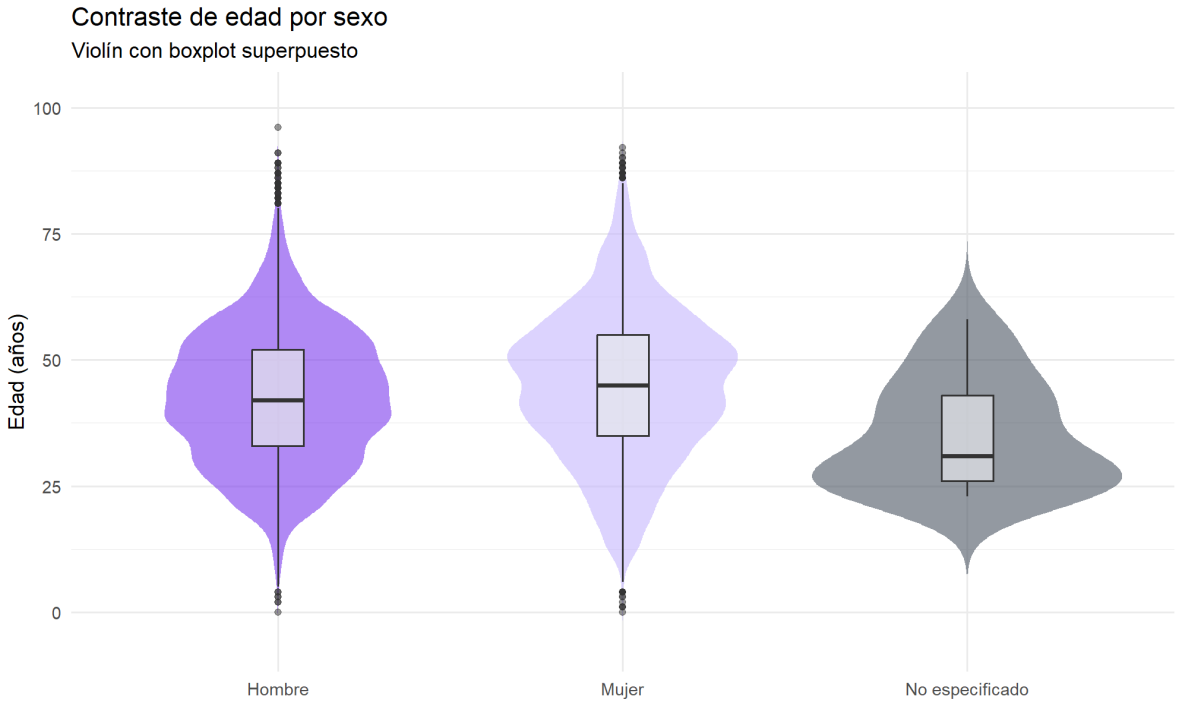
```
Warning in geom_density_ridges(alpha = 0.65, scale = 1.1, color = "#0D0C1D", :
Ignoring unknown parameters: `size`
```

```
Picking joint bandwidth of 3.05
```




```
ggplot(datos_edad_sexo_filtrado, aes(x = sexo_factor, y = Edad_limpia, fill = sexo_factor)) +
  geom_violin(trim = FALSE, alpha = 0.6, color = NA) +
  geom_boxplot(width = 0.15, fill = "#E5E7EB", alpha = 0.7, outlier.alpha = 0.5) +
  scale_fill_manual(values = paleta_sexo, guide = "none") +
  labs(
    title = "Contraste de edad por sexo",
```

```
    subtitle = "Violín con boxplot superpuesto",
    x = NULL,
    y = "Edad (años)"
  ) +
  theme_minimal(base_size = 13)
```



1.5.1 Interpretación Comparativa de la Distribución de Edad por Sexo

La combinación del gráfico de curvas de densidad y el de violín con diagrama de caja ofrece un análisis profundo y multifacético de cómo se distribuye la edad en los diferentes grupos de sexo, confirmando hallazgos y añadiendo nuevos niveles de detalle.

- **Consistente Similitud entre Hombres y Mujeres:** Ambas visualizaciones confirman de manera contundente que los perfiles de edad para hombres y mujeres son prácticamente idénticos. Las curvas de densidad muestran una forma de campana unimodal y simétrica casi superponible, y los gráficos de violín refuerzan esta idea con su forma y la posición de los diagramas de caja internos. La mediana (línea gruesa en el boxplot) y la concentración principal de casos (la parte más ancha del violín) se sitúan claramente en la adultez media para ambos grupos.

- **El Grupo "No Especificado" como una Entidad Distinta:** El hallazgo más revelador es la marcada diferencia del grupo "No especificado". El gráfico de densidad muestra que su distribución no es una campana simple, sino multimodal, con un pico dominante y muy pronunciado en una edad joven (alrededor de los 25-30 años). El gráfico de violín cuantifica esto a la perfección:
  - **Mediana Mucho Más Baja:** El diagrama de caja dentro del violín gris muestra una mediana drásticamente inferior (en torno a los 30 años) en comparación con los grupos de hombre y mujer (cercana a los 50).
  - **Forma Reveladora:** La forma del violín para este grupo no es simétrica; su base ancha en la parte inferior y su cuerpo ascendente confirman que la mayoría de los casos se concentran en edades tempranas.
- **Visualización Complementaria:** Mientras que las curvas de densidad son excelentes para mostrar la forma general y los picos de la distribución, los gráficos de violín y caja son superiores para comparar directamente las medidas estadísticas clave como la mediana y el rango intercuartílico. Juntos, no dejan lugar a dudas sobre las diferencias y similitudes.

En resumen: Los datos demuestran con claridad dos realidades distintas. Por un lado, la población principal de hombres y mujeres comparte un perfil de edad idéntico, con una incidencia centrada en la mediana edad. Por otro lado, existe un subgrupo muy pequeño pero demográficamente único, clasificado como "No especificado", que se caracteriza por una edad de afectación significativamente más temprana. Esta cohorte no es un artefacto aleatorio, sino un grupo con características propias que lo diferencian fundamentalmente del resto de la muestra.

1.6 Estancia Días por Diagnóstico y Tipo de Alta

La duración de la hospitalización está mediada tanto por la complejidad clínica del diagnóstico principal como por la vía de resolución del episodio. Para evitar que los casos ultralargos compriman la escala de análisis, se restringe la muestra al percentil 99 de `Estancia Días`; el umbral resultante se reporta más abajo.

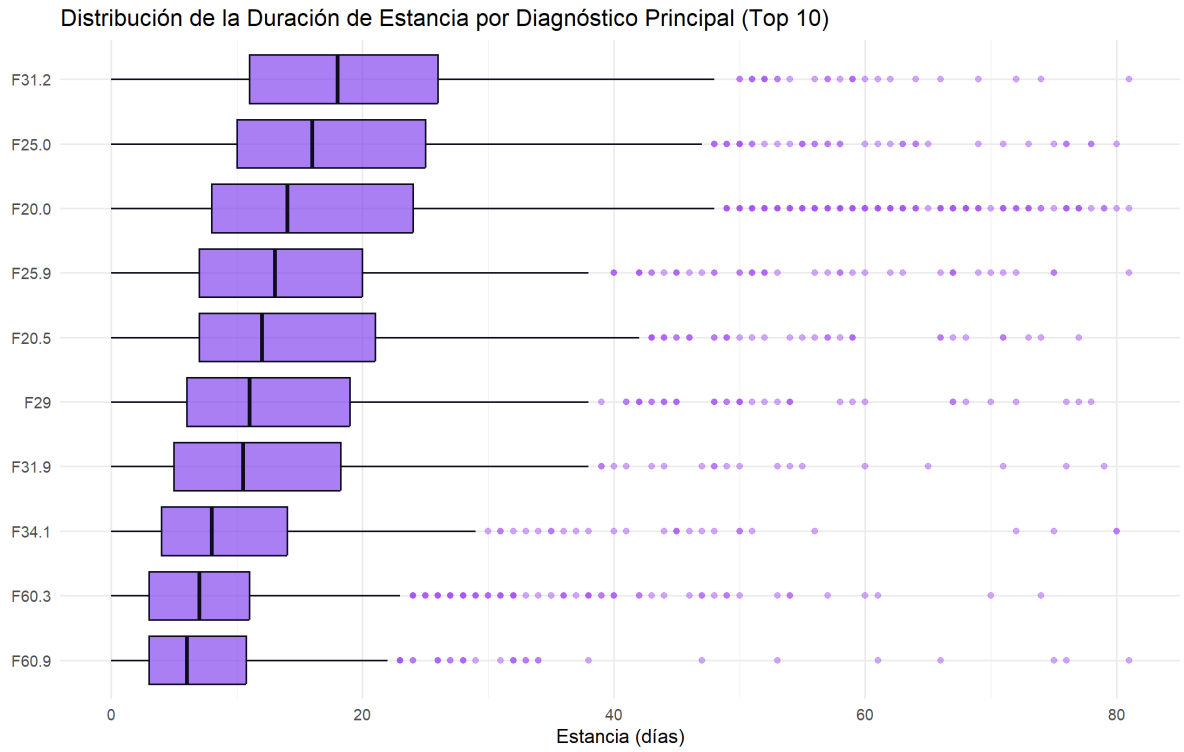
```
datos_estancia_crudos <- SaludMental %>%
  mutate(
    `Estancia Días` = readr::parse_number(as.character(`Estancia Días`)),
    `Diagnóstico Principal` = stringr::str_squish(as.character(`Diagnóstico Principal`)),
    `Tipo Alta` = stringr::str_squish(as.character(`Tipo Alta`))
  ) %>%
  filter(
    !is.na(`Estancia Días`),
```

```
    !is.na(`Diagnóstico Principal`),
    `Diagnóstico Principal` != "",
    !is.na(`Tipo Alta`),
    `Tipo Alta` != ""
  )

umbral_99_estancia <- stats::quantile(datos_estancia_crudos$`Estancia Días`, probs = 0.99, na

datos_estancia_limpios <- datos_estancia_crudos %>%
  filter(`Estancia Días` <= umbral_99_estancia)

diagnosticos_top_10 <- datos_estancia_limpios %>%
  count(`Diagnóstico Principal`, sort = TRUE) %>%
  slice_head(n = 10) %>%
  pull(`Diagnóstico Principal`)

datos_estancia_top10 <- datos_estancia_limpios %>%
  filter(`Diagnóstico Principal` %in% diagnosticos_top_10) %>%
  mutate(
    diagnostico_ordenado = forcats::fct_reorder(`Diagnóstico Principal`, `Estancia Días`, .fu
  )

resumen_diag_estancia <- datos_estancia_top10 %>%
  group_by(`Diagnóstico Principal`) %>%
  summarise(
    mediana_estancia = stats::median(`Estancia Días`),
    q3_estancia = stats::quantile(`Estancia Días`, probs = 0.75),
    n = dplyr::n(),
    .groups = "drop"
  ) %>%
  arrange(desc(mediana_estancia))

diagnostico_mayor_mediana <- resumen_diag_estancia %>% slice_max(mediana_estancia, n = 1)
diagnostico_menor_mediana <- resumen_diag_estancia %>% slice_min(mediana_estancia, n = 1)

datos_estancia_tipo_alta <- datos_estancia_limpios %>%
  mutate(
    tipo_alta_limpio = `Tipo Alta` %>%
      stringr::str_to_sentence(locale = "es") %>%
      stringr::str_replace_all("  ", " ") %>%
      stringr::str_squish()
  ) %>%
```

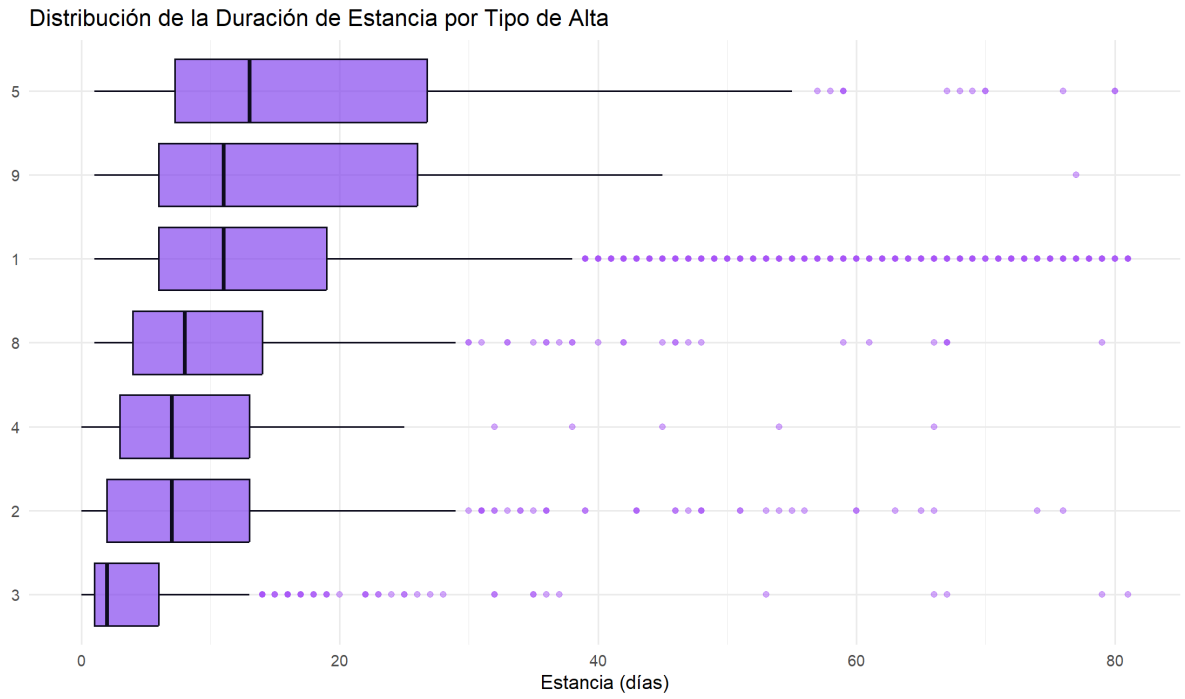
```
  filter(tipo_alta_limpio != "") %>%
  mutate(
    tipo_alta_ordenado = forcats::fct_reorder(tipo_alta_limpio, `Estancia Días`, .fun = stats
  )

resumen_tipo_alta <- datos_estancia_tipo_alta %>%
  group_by(tipo_alta_limpio) %>%
  summarise(
    mediana_estancia = stats::median(`Estancia Días`),
    n = dplyr::n(),
    .groups = "drop"
  ) %>%
  arrange(desc(mediana_estancia))

tipo_alta_mayor_mediana <- resumen_tipo_alta %>% slice_max(mediana_estancia, n = 1)
tipo_alta_menor_mediana <- resumen_tipo_alta %>% slice_min(mediana_estancia, n = 1)

porcentaje_retencion_estancia <- sprintf(
  "%.1f",
  100 * nrow(datos_estancia_limpios) / nrow(datos_estancia_crudos)
)
```

En total, la depuración mantiene 99.0% de los episodios con información válida. El recorte al percentil 99 equivale a 81 días de estancia, valor a partir del cual los casos son muy excepcionales y se analizan aparte.

```
ggplot(datos_estancia_top10, aes(x = `Estancia Días`, y = diagnostico_ordenado)) +
  geom_boxplot(
    fill = "#7C3AED",
    alpha = 0.65,
    outlier.color = "#A855F7",
    outlier.alpha = 0.5,
    color = "#0D0C1D"
  ) +
  labs(
    title = "Distribución de la Duración de Estancia por Diagnóstico Principal (Top 10)",
    x = "Estancia (días)",
    y = NULL
  ) +
  theme_minimal(base_size = 13)
```




```r
ggplot(datos_estancia_tipo_alta, aes(x = `Estancia Días`, y = tipo_alta_ordenado)) +
  geom_boxplot(
    fill = "#7C3AED",
    alpha = 0.65,
    outlier.color = "#A855F7",
    outlier.alpha = 0.5,
    color = "#0D0C1D"
  ) +
  labs(
    title = "Distribución de la Duración de Estancia por Tipo de Alta",
    x = "Estancia (días)",
    y = NULL
  ) +
  theme_minimal(base_size = 13)
```



1.6.1 Interpretación de la Duración de Estancia por Diagnóstico Principal

Este gráfico de diagramas de caja (boxplots) ofrece una comparación detallada de la duración de la estancia hospitalaria para los 10 diagnósticos principales, revelando un patrón consistente y muy informativo.

- **Notable Homogeneidad en la Estancia Típica:** La observación más sorprendente es la gran similitud en la distribución de las estancias para la mayoría de los pacientes, independientemente del diagnóstico. Las "cajas" (que representan el 50% central de los casos) son muy parecidas en longitud y posición para casi todos los códigos. Esto indica que la duración de una hospitalización "estándar" o "típica" es muy consistente, con medianas que se sitúan generalmente entre los 10 y 15 días.
- **La Universalidad de las Estancias Atípicas:** El hallazgo más crítico es la presencia de una larga cola de valores atípicos (outliers), representados por los puntos a la derecha de cada diagrama. Este patrón no es exclusivo de un diagnóstico, sino una característica **universal** en todos los grupos. Esto significa que en cada una de las patologías más frecuentes, existe un subgrupo de pacientes cuyas estancias se prolongan de manera excepcional, superando con creces la norma.
- **Sutiles Diferencias entre Diagnósticos:** Aunque el patrón general es homogéneo, se pueden apreciar ligeras diferencias. Por ejemplo, los diagnósticos **F31.2 (Trastorno afectivo bipolar, episodio maníaco con síntomas psicóticos)** y **F25.0 (Trastorno**

esquizoafectivo, tipo maníaco) parecen tener las cajas ligeramente desplazadas hacia la derecha, sugiriendo que sus estancias típicas tienden a ser un poco más largas en comparación con otras, como el **F60.3 (Trastorno de inestabilidad emocional de la personalidad)**.

En resumen: Si bien la gran mayoría de las hospitalizaciones siguen un curso predecible y de duración similar (entre 1 y 3 semanas) sin importar el diagnóstico principal, el verdadero desafío para la gestión de recursos y la planificación clínica reside en los casos atípicos. La existencia de pacientes con estancias extremadamente largas es un fenómeno sistémico que afecta a todas las patologías principales, y no un problema aislado de una condición específica. Identificar los factores que predicen estas estancias prolongadas es, por tanto, mucho más crucial que analizar las pequeñas diferencias en las estancias típicas.