EDA Salud Mental

Table of contents

0.1 Librerías

```
library(tidyverse)
```

```
-- 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
```

```r
library(readxl)
library(visdat)
library(GGally)

# Opciones de gráficos para acelerar el render a PDF
knitr::opts_chunk$set(dev = "jpeg", dpi = 120, dev.args = list(quality = 80))
```

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 'OUT94ZZ'

Warning: Expecting logical in CY1831 / R1831C103: got 'OUT24ZZ'

Warning: Expecting logical in CZ1831 / R1831C104: got 'OUT74ZZ'

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'
```

```r
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
```

```r
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>, ...
```

Comentarios sobre la carga

- El dataset tiene 21210 filas y 111 columnas.

- Durante la importación inicial del conjunto de datos, se mostraron advertencias del tipo Expecting logical … got '0UT94ZZ'. Estos avisos se produjeron porque R infirió incorrectamente que algunas columnas contenían datos lógicos (verdadero/falso) basándose en sus primeras filas, para luego encontrar códigos de texto que no coincidían con esa suposición. Para resolver esto y prevenir cualquier pérdida de información, se especificó que estas columnas se importaran como texto, garantizando así que todos los valores originales fueran conservados íntegramente para el análisis.

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
```

Interpretación

El conjunto de datos contiene 21210 filas y 111 columnas. La distribución de tipos de datos es: character, logical, numeric, POSIXct.

Esta distribución inicial, sin embargo, enmascara la verdadera naturaleza de las variables. Se han identificado 26 columnas (logical) que están completamente vacías y serán eliminadas. Adicionalmente, numerosas variables leídas como numéricas (Sexo) o texto (País Nacimiento) son en realidad categorías codificadas que se convertirán a factores. Finalmente, columnas de fecha mal interpretadas (Fecha de Fin Contacto) serán corregidas. Estos pasos de limpieza son esenciales antes de proceder con el análisis visual y estadístico.

0.4 Valores faltantes (nulos)

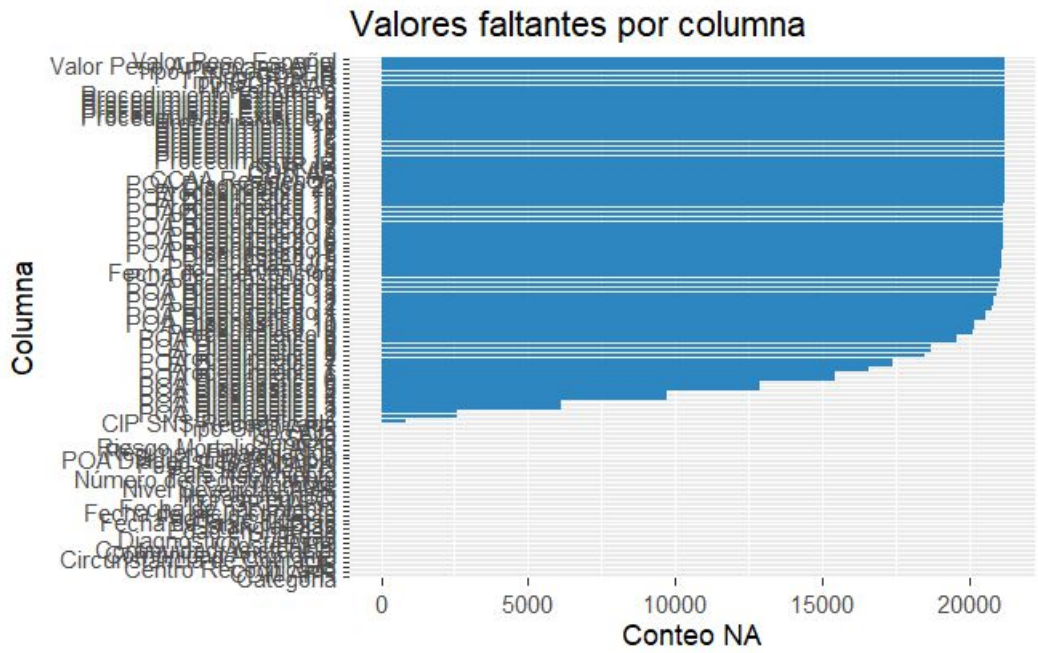
```r
na_por_col <- SaludMental %>%
  summarise(across(everything(), ~ sum(is.na(.)))) %>%
  pivot_longer(everything(), names_to = "columna", values_to = "na") %>%
  arrange(desc(na)) %>%
  mutate(pct_na = 100 * na / num_filas)

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

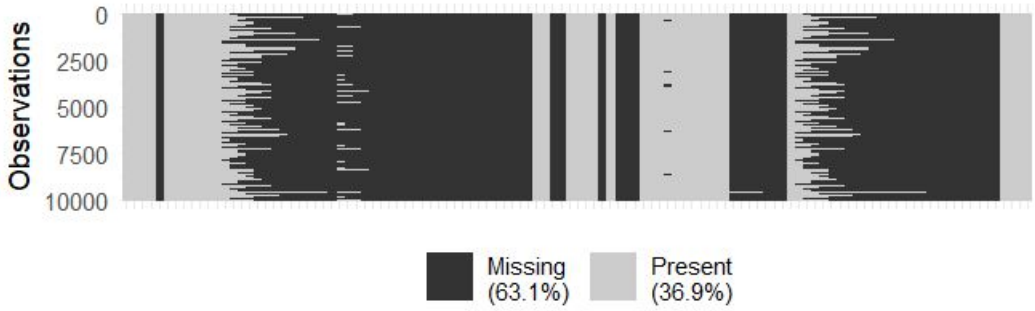
```
# A tibble: 111 x 3
   columna             na pct_na
   <chr>            <int>  <dbl>
 1 CCAA Residencia  21210    100
 2 Procedimiento 12 21210    100
 3 Procedimiento 13 21210    100
 4 Procedimiento 14 21210    100
 5 Procedimiento 15 21210    100
```

```
 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
```

```
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")
```




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



Interpretación

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

1. Columnas Completamente Vacías (100% NA) El hallazgo más evidente, visible tanto en el gráfico de barras como en la tabla, es la existencia de un gran número de columnas (más de 25) que están completamente vacías.

Evidencia: Columnas como CCAA Residencia, Procedimiento 12 hasta Procedimiento 20, GDR IR, y muchas otras, tienen 21,210 valores faltantes, lo que corresponde al 100% de las filas.

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>, ...
```

Interpretación

El análisis de duplicados ha confirmado que el conjunto de datos no contiene ninguna fila repetida. El conteo de Filas duplicadas es 0, lo que representa el 0% del total de las observaciones.

Esto es un indicador excelente de la calidad de los datos, ya que significa que cada registro en el dataset es único y representa una observación distinta. Por lo tanto, no es necesario realizar ninguna acción de limpieza para eliminar duplicados, y podemos proceder con el análisis con la confianza de que no estamos sobrecontando ninguna entrada.

0.6 Variables numéricas: resúmenes y gráficos

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

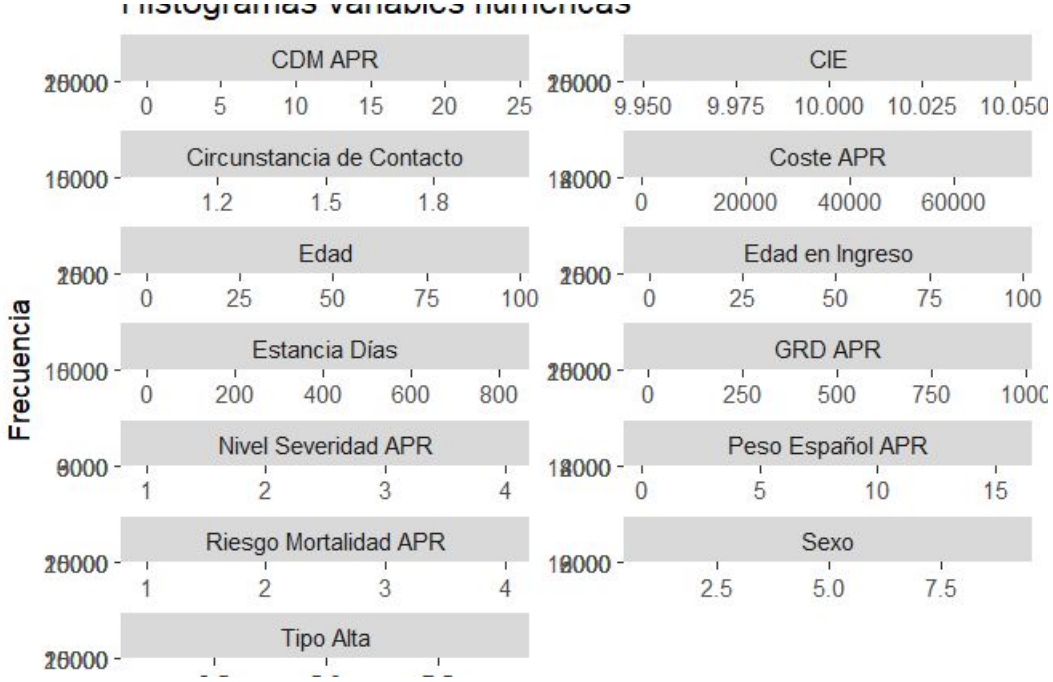
```
Variables numéricas: 13
```

```
summary(numericas)
```

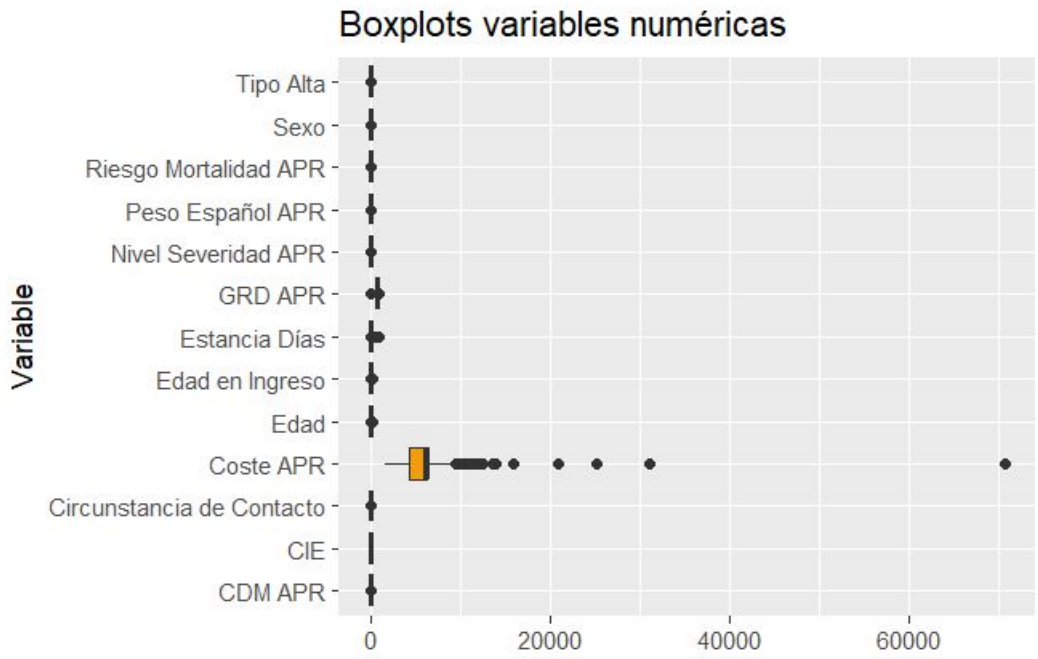
```
      Sexo       Circunstancia de Contacto   Tipo Alta     Estancia Días
 Min.   :1.000   Min.   :1.000             Min.   :1.000   Min.   :  0.00
 1st Qu.:1.000   1st Qu.:1.000             1st Qu.:1.000   1st Qu.:  5.00
 Median :1.000   Median :1.000             Median :1.000   Median : 11.00
 Mean   :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")
```

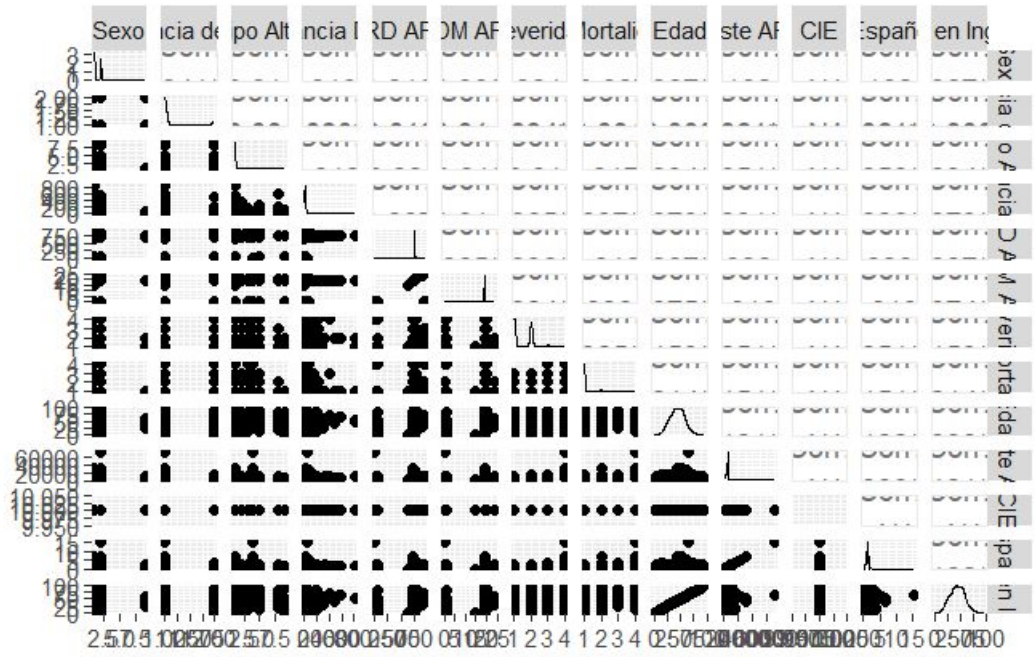



```r
numericas %>%
pivot_longer(everything(), names_to = "variable", values_to = "valor") %>%
ggplot(aes(x = variable, y = valor)) +
geom_boxplot(fill = "#F39C12") +
coord_flip() +
labs(title = "Boxplots variables numéricas", x = "Variable", y = NULL)
```




```
GGally::ggpairs(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 = "")
```

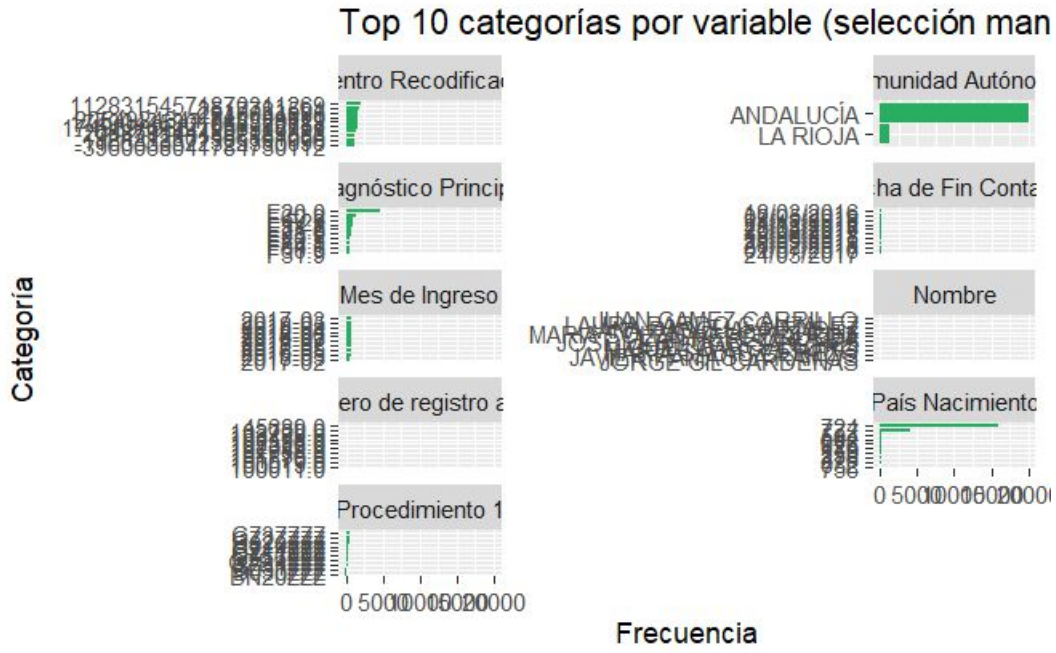
```
Variables: Comunidad Autónoma, Nombre, Fecha de Fin Contacto, Diagnóstico Principal, Procedi
```

```
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")

  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)")
    )
```

```
} else {
  cat("Ninguna de las variables seleccionadas está presente en el dataset.\n")
}
```



Este análisis de las 9 variables categóricas seleccionadas manualmente ofrece una visión clara de la composición del dataset, revelando tanto patrones importantes como áreas que requieren atención en la limpieza de datos.

Interpretación de los Gráficos Los resultados se pueden agrupar en cuatro categorías principales de hallazgos:

1. Fuerte Sesgo Geográfico y Demográfico Comunidad Autónoma: El hallazgo más impactante es el desequilibrio extremo en esta variable. La inmensa mayoría de los registros pertenecen a ANDALUCÍA, con una representación mínima de otras comunidades como LA RIOJA. Esto es crucial: cualquier conclusión extraída de este análisis será representativa de Andalucía, pero no puede generalizarse al resto de España.

País Nacimiento: Se observa un patrón similar. El código "724" (que corresponde a España) es abrumadoramente dominante. El código "ZZZ" probablemente representa un valor desconocido o no especificado. Esto indica que el dataset se centra casi exclusivamente en pacientes nacidos en España.

2. Información Clínica Relevante Diagnóstico Principal: Este gráfico es muy informativo. Muestra que los diagnósticos más frecuentes son códigos CIE-10 del capítulo de

trastornos mentales y del comportamiento (códigos 'F'). Destacan el F20.9 (Esquizofrenia no especificada), F20.0 (Esquizofrenia paranoide) y F31.9 (Trastorno afectivo bipolar no especificado), dándonos una primera idea de las patologías más comunes en esta cohorte de pacientes.

Procedimiento 1: Muestra los códigos de los procedimientos iniciales más comunes. El código GZZZZZZ es el más frecuente, lo cual en sistemas de codificación como ICD-10-PCS a menudo significa "Sin procedimiento" o un código genérico. Los otros códigos (B02ZZZZ, GZ14ZZZ, etc.) apuntan a intervenciones específicas que valdría la pena investigar.

3. Identificadores de Alta Cardinalidad Nombre, Número de registro anual y Centro Recodificado: Estas variables se caracterizan por tener una alta cardinalidad, lo que significa que tienen muchos valores únicos. Como se ve en los gráficos, incluso las 10 categorías más frecuentes tienen una frecuencia muy baja. Esto es esperado, ya que son identificadores (de persona, de registro y de centro, respectivamente) y no variables categóricas para análisis de grupos. Su utilidad principal es identificar registros individuales, no para agregación.

4. Variables Temporales Mal Tipificadas Mes de Ingreso y Fecha de Fin Contacto: Ambas variables son de naturaleza temporal, pero están siendo tratadas como texto (character). El gráfico de Mes de Ingreso nos permite ver los meses con más admisiones (como 2017-03), pero para un análisis de tendencias adecuado, estas columnas deben ser convertidas a un formato de fecha/hora. Tratar cada fecha única como una categoría distinta, como en Fecha de Fin Contacto, no es analíticamente útil.

Conclusión General El análisis de estas variables categóricas ha sido un éxito. Ha revelado la limitación geográfica clave del estudio (centrado en Andalucía), ha proporcionado un perfil clínico inicial de los pacientes, ha confirmado la naturaleza de las variables de identificación y ha señalado la necesidad crítica de corregir los tipos de datos de las variables de fecha.

0.8 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
```

```
  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%)
```

Interpretación del Análisis de Outliers La tabla identifica correctamente los valores que caen fuera del rango intercuartílico (IQR), pero es fundamental separar los resultados en dos grupos para una interpretación correcta.

1. Falsos Positivos: Outliers en Variables Categóricas La mayoría de las variables que encabezan la lista con el mayor porcentaje de outliers (Circunstancia de Contacto, Tipo Alta, GRD APR, Riesgo Mortalidad, etc.) son, como identificamos anteriormente, variables categóricas codificadas como números.

¿Por qué ocurre esto? El método del IQR está diseñado para datos continuos. Cuando se aplica a datos categóricos con pocos valores (ej. 1 y 2), el rango intercuartílico (iqr) puede ser 0 o 1.

En el caso de Circunstancia de Contacto, la mayoría de los valores son '1'. Esto hace que tanto el primer cuartil (q1) como el tercer cuartil (q3) sean '1', resultando en un iqr de 0. La fórmula entonces considera cualquier valor que no sea '1' (como el '2') como un outlier.

Conclusión: Los 10.5% de "outliers" en Circunstancia de Contacto no son errores ni valores extremos, son simplemente las observaciones que pertenecen a la segunda categoría menos frecuente. Lo mismo ocurre con las otras variables de esta lista. Estos no son outliers que deban ser eliminados, sino una confirmación de que estas variables deben ser tratadas como factores.

2. Outliers Reales y Significativos En contraste, los outliers identificados en las variables verdaderamente cuantitativas son informativos y esperados.

Estancia Días (5.8% outliers): Estos valores representan a los pacientes con estancias hospitalarias inusualmente largas. Son datos muy valiosos que podrían indicar casos más complejos, complicaciones o perfiles de pacientes específicos. No deben ser eliminados a la ligera, sino estudiados en detalle.

Coste APR (0.9% outliers) y Peso Español APR (0.8% outliers): Estos son los casos de alto coste. Como vimos en los histogramas, la distribución de costes está muy sesgada. Estos outliers representan los tratamientos más caros y son de gran interés para el análisis de gestión y eficiencia.

Edad y Edad en Ingreso (0.7% outliers): Corresponden a los pacientes muy jóvenes o muy mayores en comparación con el grueso de la población del dataset.

Resumen General El análisis de outliers ha sido extremadamente útil, no tanto para encontrar errores, sino para confirmar la naturaleza de nuestras variables.

Ha reforzado la necesidad de convertir las variables categóricas a factores para evitar interpretaciones incorrectas.

Ha validado la presencia de valores extremos significativos en las métricas clave como Estancia Días y Coste APR, los cuales son el foco de muchos análisis clínicos y económicos.