Análisis Exploratorio de SaludMental

Castores Afanosos

Este análisis tiene como objetivo principal proporcionar una visión estadística elemental y una evaluación de la calidad de los datos contenidos en el dataset de salud mental, abarcando la identificación de tipos de variables, la cuantificación de datos faltantes (nulos o desconocidos) y la detección preliminar de valores atípicos (outliers).

```
SaludMental <- read_excel("SaludMental.xls")
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>, ...
```

Análisis descriptivo

Estructura y Tipos de Datos

En primer lugar, la estructura del conjunto de datos es revisada para verificar que se han inferido correctamente el tipo de cada variable.

- **Variables categóricas**. Columnas como la *comunidad*, el *sexo*, *circunstancia de contacto* o *diagnóstico,* las cuales, transformaremos en factores más adelante.
- **Variables numéricas.** Variables como *Estancia Días* o *Edad* fueron confirmadas como de tipo entero (`int`), double(`dbl`) o numérico (`num`), permitiendo el cálculo de medidas de tendencia central y dispersión.
- **Tipos de Datos Desconocidos.** Se identificaron algunas variables como *CCAA Residencia* o *Reingreso* en las que no hay datos, por lo que no se tiene información de tipo, siendo interesante investigar si se puede reconstruir o eliminarlas.

Análisis Estadístico Elemental y Outliers

```
SaludMental |>
  select(c(1:10)) |>
  summary()
```

```
##  Comunidad Autónoma    Nombre          Fecha de nacimiento
##  Length:21210       Length:21210       Min.   :1921-03-07 00:00:00
##  Class :character   Class :character   1st Qu.:1963-09-23 00:00:00
##  Mode  :character   Mode  :character   Median :1973-03-03 00:00:00
##                                        Mean   :1973-05-14 00:29:19
##                                        3rd Qu.:1983-04-20 00:00:00
##                                        Max.   :2018-09-19 00:00:00
##       Sexo       CCAA Residencia Fecha de Ingreso
##  Min.   :1.000   Mode:logical    Min.   :2016-01-01 00:00:00
##  1st Qu.:1.000   NA's:21210      1st Qu.:2016-10-02 00:00:00
##  Median :1.000                   Median :2017-06-30 00:00:00
##  Mean   :1.451                   Mean   :2017-07-04 15:19:52
##  3rd Qu.:2.000                   3rd Qu.:2018-04-13 00:00:00
##  Max.   :9.000                   Max.   :2018-12-31 00:00:00
##  Circunstancia de Contacto Fecha de Fin Contacto   Tipo Alta
##  Min.   :1.000             Length:21210          Min.   :1.000
##  1st Qu.:1.000             Class :character      1st Qu.:1.000
##  Median :1.000             Mode  :character      Median :1.000
##  Mean   :1.105                                   Mean   :1.263
##  3rd Qu.:1.000                                   3rd Qu.:1.000
##  Max.   :2.000                                   Max.   :9.000
##  Estancia Días
##  Min.   :  0.00
##  1st Qu.:  5.00
##  Median : 11.00
##  Mean   : 15.46
##  3rd Qu.: 19.00
##  Max.   :814.00
```

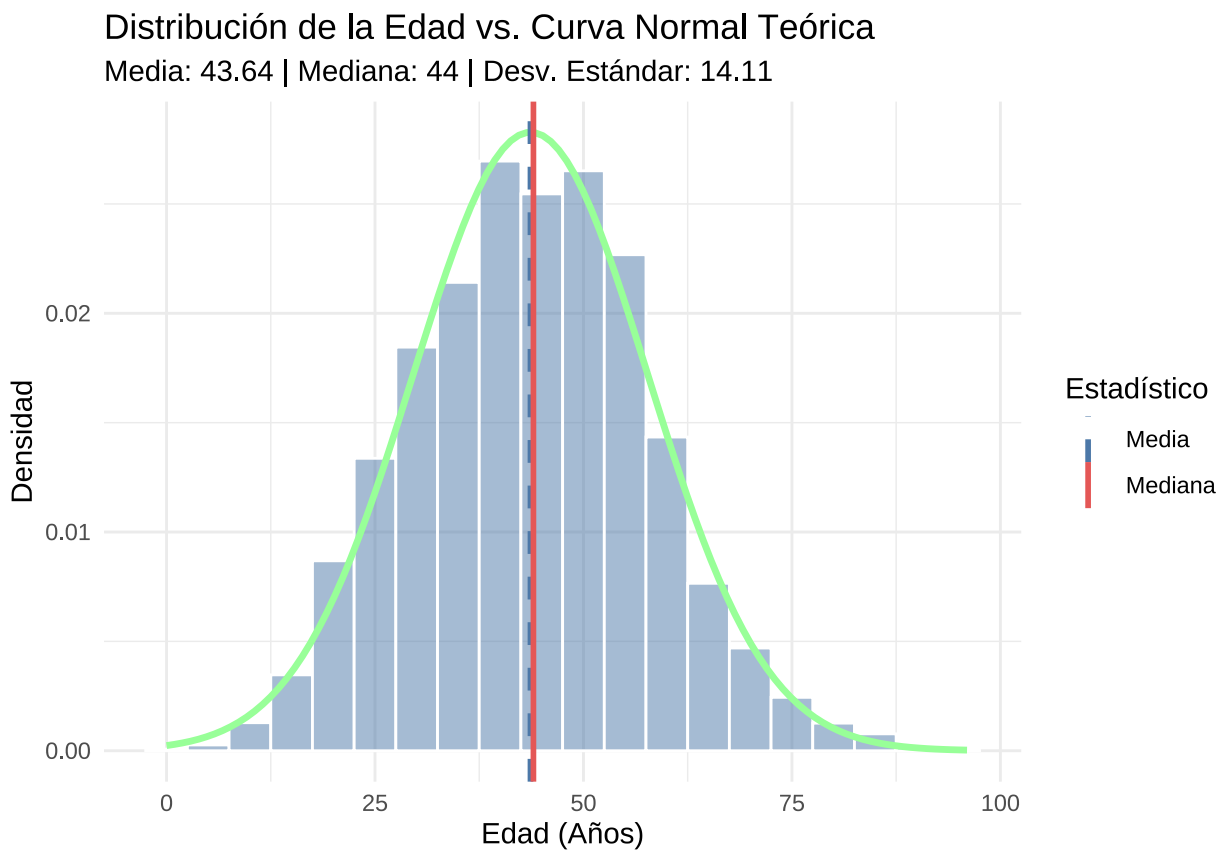
Observamos que en algunas variables claves como la *Edad*, en la que la media y la mediana son muy cercanas, sugiriendo que sigue una distribución normal o muy ligeramente sesgada.

```
summary(SaludMental$Edad)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   34.00   44.00   43.64   53.00   96.00
```

```
# Calcular Media, Mediana y Desviación Estándar (excluyendo NA)
media_edad <- mean(SaludMental$Edad, na.rm = TRUE)
sd_edad <- sd(SaludMental$Edad, na.rm = TRUE)
mediana_edad <- median(SaludMental$Edad, na.rm = TRUE)

# Generar el gráfico con la curva Gaussiana (verde)
SaludMental |>
  ggplot(aes(x = Edad)) +
```

```r
# Histograma (Muestra la frecuencia real de los datos)
geom_histogram(
  aes(y = after_stat(density)),
  binwidth = 5,
  fill = "#4C78A8",
  color = "white",
  alpha = 0.5
) +
# Curva GAUSSIANA
geom_function(
  fun = dnorm, # La función de densidad normal
  args = list(mean = media_edad, sd = sd_edad),
  color = "#98FF98",
  linewidth = 1.2
) +

# Marcas de Media y Mediana
geom_vline(aes(xintercept = media_edad, color = "Media"),
           linetype = "dashed", linewidth = 1) +
geom_vline(aes(xintercept = mediana_edad, color = "Mediana"),
           linetype = "solid", linewidth = 1) +

# Etiquetas y tema
scale_color_manual(name = "Estadístico",
                   values = c("Media" = "#4C78A8", "Mediana" = "#E45756")) +
labs(
  title = "Distribución de la Edad vs. Curva Normal Teórica",
  subtitle = paste0("Media: ", round(media_edad, 2),
                    " | Mediana: ", round(mediana_edad, 2),
                    " | Desv. Estándar: ", round(sd_edad, 2)),
  x = "Edad (Años)",
  y = "Densidad"
) +
theme_minimal()
```



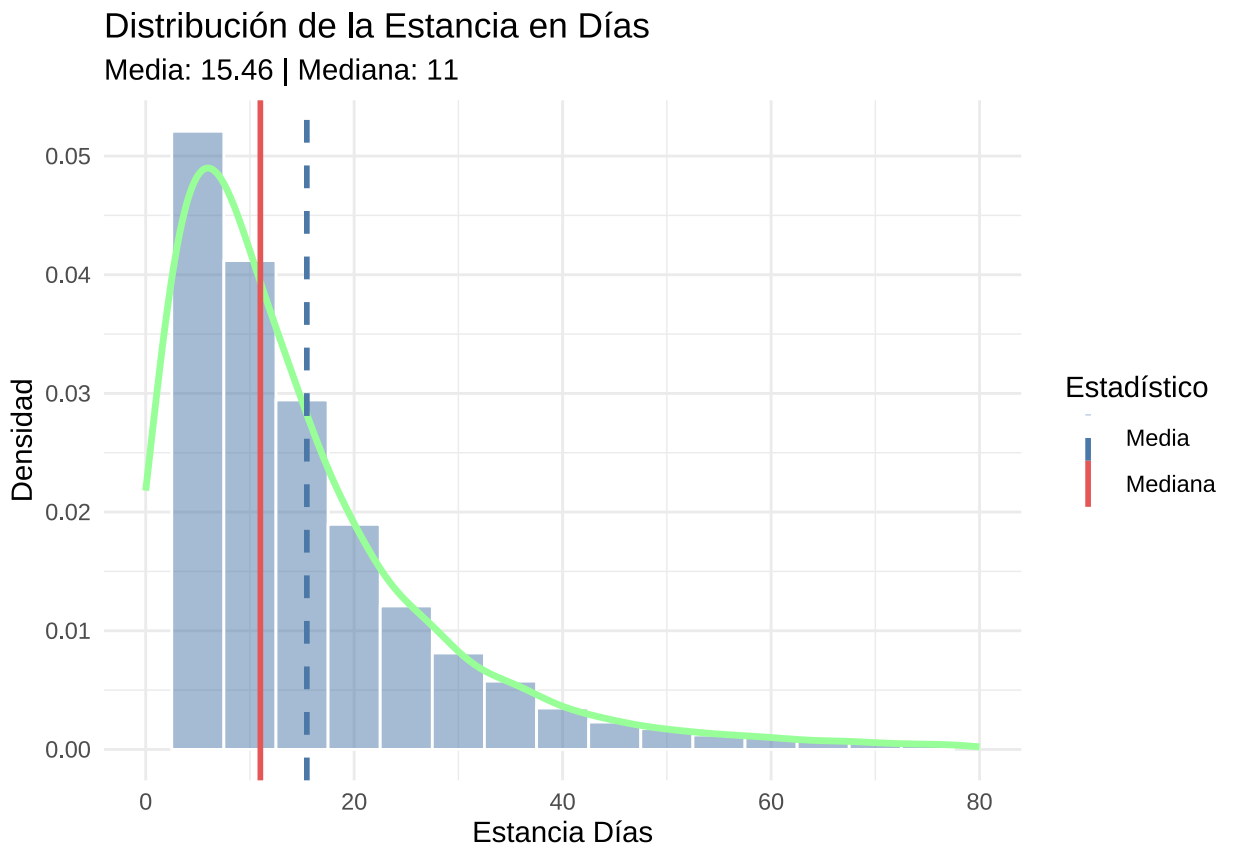
Por otro lado, en variables como *Estancia Dias,* el valor máximo (814) está muy por encima de la media (15,46), lo que confirma la presencia de outliers extremos los cuales, elevan el valor de la media con relación a la mediana (11), dando lugar a una distribución sesgada.

```
summary(SaludMental$`Estancia Días`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    5.00   11.00   15.46   19.00  814.00
```

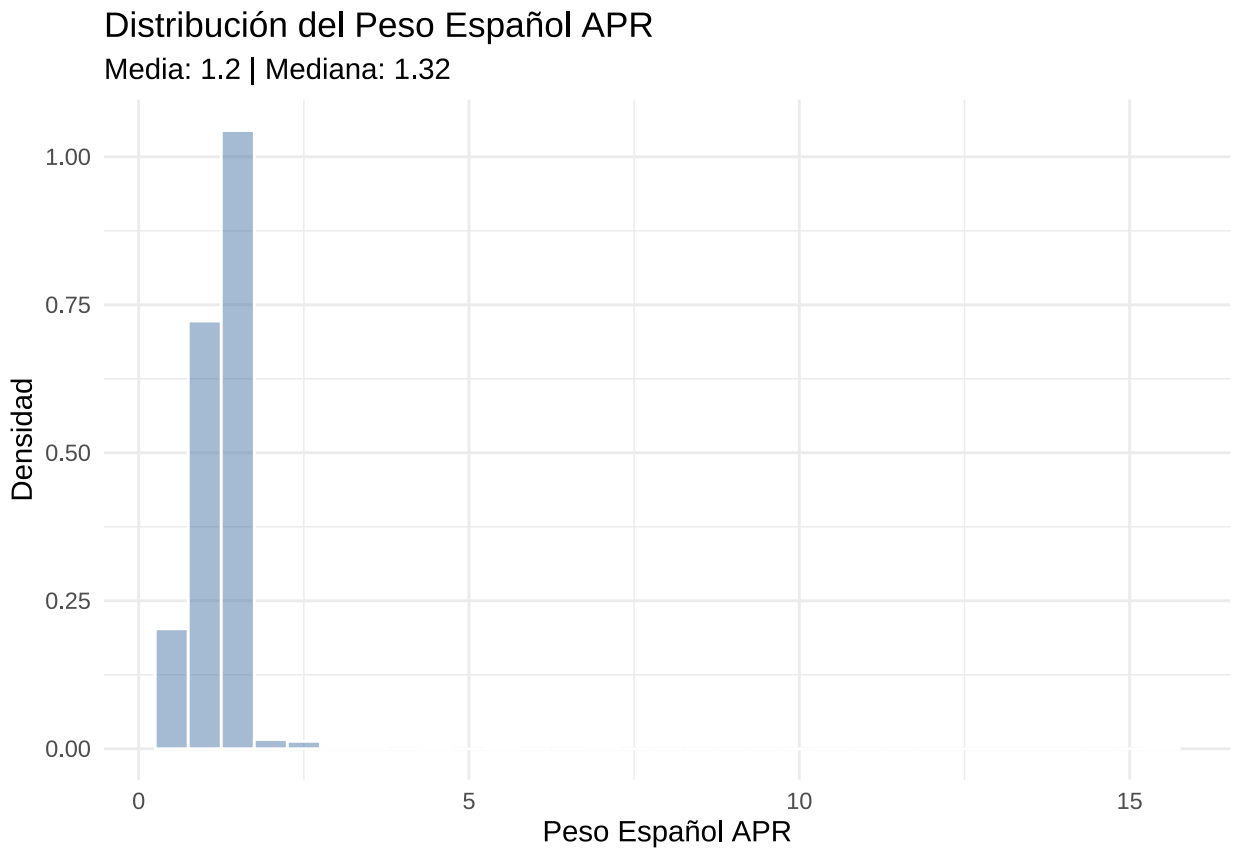
```
# Calcular Media, Mediana de Estancia Días (excluyendo NA)
media_estancia <- mean(SaludMental$`Estancia Días`, na.rm = TRUE)
mediana_estancia <- median(SaludMental$`Estancia Días`, na.rm = TRUE)

# Generar el gráfico de distribución
SaludMental |>
  ggplot(aes(x = `Estancia Días`)) +
  # Histograma (Muestra la frecuencia real de los datos)
  geom_histogram(
    aes(y = after_stat(density)),
    binwidth = 5, # Agrupa los días en intervalos de 5
    fill = "#4C78A8",
    color = "white",
    alpha = 0.5
  ) +
  # Curva de Densidad
```

```r
  geom_density(linewidth = 1.2, color = "#98FF98", adjust = 2) +

  # Marcas de Media y Mediana
  geom_vline(aes(xintercept = media_estancia, color = "Media"),
             linetype = "dashed", linewidth = 1) +
  geom_vline(aes(xintercept = mediana_estancia, color = "Mediana"),
             linetype = "solid", linewidth = 1) +

  # Limitamos el eje X para evitar que el outlier de 814 comprima el gráfico
  xlim(0, 80) +

  scale_color_manual(name = "Estadístico",
                     values = c("Media" = "#4C78A8", "Mediana" = "#E45756")) +
  labs(
    title = "Distribución de la Estancia en Días",
    subtitle = paste0("Media: ", round(media_estancia, 2),
                      " | Mediana: ", round(mediana_estancia, 2)),
    x = "Estancia Días",
    y = "Densidad"
  ) +
  theme_minimal()
```



De la misma forma, la variable *Peso Español APR* tiene un sutil sesgo negativo, ya que al igual que la variable anterior, contiene potenciales outliers.

```r
# Calcular Media, Mediana y Desviación Estándar de Estancia Días (excluyendo NA)
media_pesoEsp <- mean(SaludMental$`Peso Español APR`, na.rm = TRUE)
mediana_pesoEsp <- median(SaludMental$`Peso Español APR`, na.rm = TRUE)

#  Generar el gráfico de distribución
SaludMental |>
  ggplot(aes(x = `Peso Español APR`)) +
  #  Histograma (Muestra la frecuencia real de los datos)
  geom_histogram(
    aes(y = after_stat(density)),
    binwidth = 0.5, # Agrupa los días en intervalos de 0.5
    fill = "#4C78A8",
    color = "white",
    alpha = 0.5
  ) +

  # Etiquetas y tema
  scale_color_manual(name = "Estadístico",
                     values = c("Media" = "#4C78A8", "Mediana" = "#E45756")) +
  labs(
    title = "Distribución del Peso Español APR",
    subtitle = paste0("Media: ", round(media_pesoEsp, 2),
                      " | Mediana: ", round(mediana_pesoEsp, 2)),
    x = "Peso Español APR",
    y = "Densidad"
  ) +
  theme_minimal()
```



Análisis de Variables Categóricas

Como se ha mencionado, en primer lugar transformaremos las variables principales a factor, re-etiquetando las que sea necesario de acuerdo con las especificaciones dadas.

```
SaludMental <- SaludMental |>
  mutate(
    # **SEXO:** 1. Varón / 2. Mujer / 3. Indeterminado / 9. No especificado
    Sexo = factor(
      Sexo,
      levels = c(1, 2, 3, 9),
      labels = c("Varón", "Mujer", "Indeterminado", "No especificado")
    ),

    # **Tipo Alta:** 1. Domicilio / 2. Traslado Hospital / 3. Alta voluntaria
    #  4. Éxitus / 5. Traslado Sociosanitario / 9. Otros/Desconocido
    `Tipo Alta` = factor(
      `Tipo Alta`,
      levels = c(1, 2, 3, 4, 5, 9),
      labels = c("Domicilio", "Traslado Otro Hospital", "Alta Voluntaria",
                 "Éxitus", "Traslado Sociosanitario", "Otros/Desconocido")
    ),

    # **Régimen de financiación:** 1 a 9 según diccionario
    `Régimen Financiación` = factor(
```

```
      `Régimen Financiación`,
      levels = c(1, 2, 3, 4, 5, 6, 7, 8, 9),
      labels = c("Seguridad Social", "Corporaciones Locales/Cabildos",
                 "Mutuas de Asistencia", "Accidentes de Trabajo",
                 "Accidentes de Tráfico", "Privado", "Financiación Mixta",
                 "Otros", "Desconocido")
    ),

    # **Circunstancia de Contacto (Asumiendo que es Tipo de Ingreso):** 1. Urgente /
    #2. Programado / 9. Otros/Desconocido
    `Circunstancia de Contacto` = factor(
      `Circunstancia de Contacto`,
      levels = c(1, 2, 9), # Asumiendo que 9 es el código de Otros/Desconocido si existe.
      labels = c("Urgente", "Programado", "Otros/Desconocido")
    )
  )
```

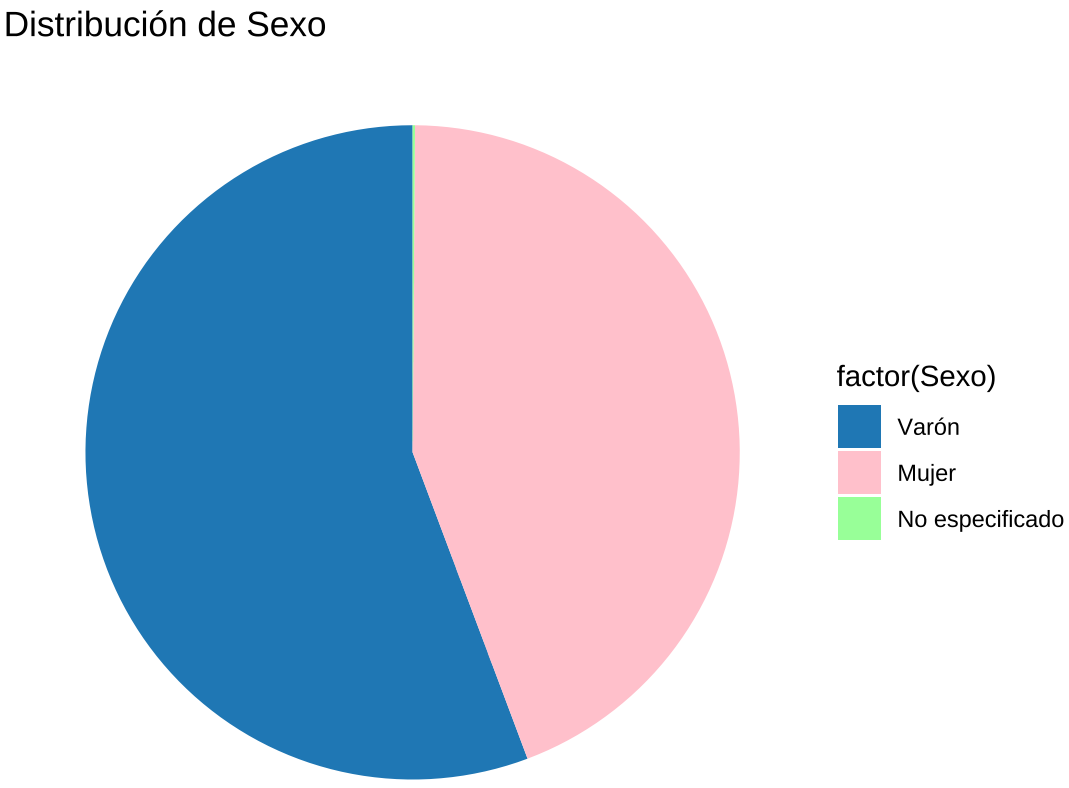
Una vez re-etiquetadas, se convierte a factor las que quedan.

```
 # Definir una lista de las variables categóricas
variables_categoricas <- c("Comunidad Autónoma", "Categoría", "Servicio",
                           "Nivel Severidad APR", "Riesgo Mortalidad APR",
                           "Tipo GRD APR", "Mes de Ingreso")

SaludMental <- SaludMental |>
  mutate(
    across(
      .cols = all_of(variables_categoricas),
      .fns = as.factor
    )
  )
# Aplicar la conversión a factor usando mutate(across())
SaludMental <- SaludMental |>
  mutate(
    # 'across' permite aplicar una función a múltiples columnas
    across(
      .cols = all_of(variables_categoricas), # Selecciona las columnas de la lista
      .fns = as.factor                       # La función a aplicar (convertir a factor)
    )
  )


str(SaludMental[variables_categoricas])
```

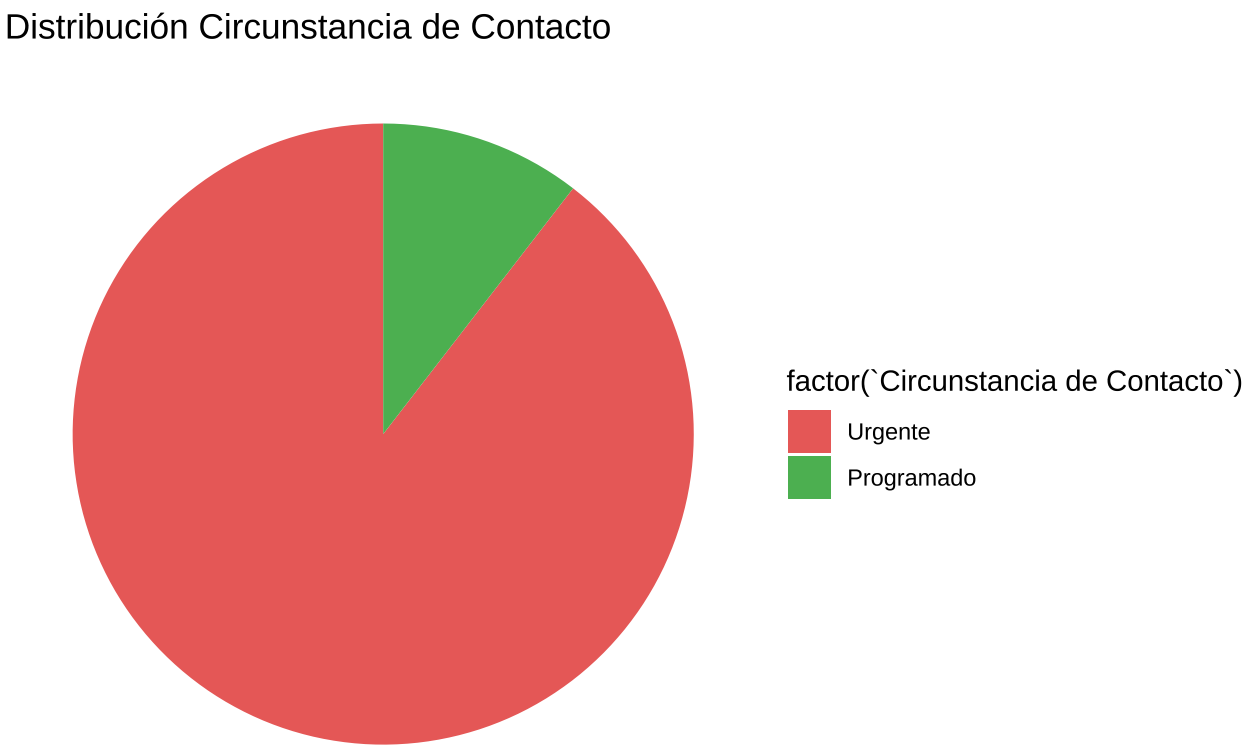
```
## tibble [21,210 x 7] (S3: tbl_df/tbl/data.frame)
##  $ Comunidad Autónoma   : Factor w/ 2 levels "ANDALUCÍA","LA RIOJA": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Categoría            : Factor w/ 7 levels "Esquizofrenia, trastornos esquizotípicos y trastornos d
##  $ Servicio             : Factor w/ 30 levels "ACV","ALG","CAR",..: 26 3 26 26 26 26 15 26 26 26 ...
##  $ Nivel Severidad APR  : Factor w/ 4 levels "1","2","3","4": 2 1 2 1 1 1 2 2 1 1 ...
##  $ Riesgo Mortalidad APR: Factor w/ 4 levels "1","2","3","4": 1 2 1 2 1 1 1 2 1 1 ...
##  $ Tipo GRD APR         : Factor w/ 2 levels "M","Q": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Mes de Ingreso       : Factor w/ 36 levels "2016-01","2016-02",..: 1 1 1 1 1 1 1 1 1 1 ...
```

A continuación, se representarán algunas

```r
SaludMental |>
   ggplot(aes(x = "", fill = factor(Sexo))) +
  geom_bar(stat = "count", width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Distribución de Sexo", x = NULL, y = NULL) +
  theme_void() +
  scale_fill_manual(values = c("#1F77B4", "#FFC0CB", "#98FF98"))
```




```r
SaludMental |>
   ggplot(aes(x = "", fill = factor(`Circunstancia de Contacto`))) +
  geom_bar(stat = "count", width = 1) +
  coord_polar(theta = "y") +
  labs(title = " Distribución Circunstancia de Contacto", x = NULL, y = NULL) +
  theme_void() +
  scale_fill_manual(values = c("#E45756","#4CAF50"))
```



Estudio de Valores Desconocidos (NA)

En este estudio veremos el porcentaje de valores NA del dataset.

Se visualizarán las variables de tipo numérico y factor (categórico)

```
SaludMental |>
  select(
    where(is.numeric), # Selecciona todas las columnas de tipo numérico/entero
    where(is.factor)   # Selecciona todas las columnas de tipo factor
  ) |>
  skim()
```

Table 1: Data summary

Name	select(SaludMental, where…
Number of rows	21210
Number of columns	19
_____	
Column type frequency:	
factor	11
numeric	8
_____	
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
Comunidad Autónoma	0	1.00	FALSE	2	AND: 20034, LA : 1176
Sexo	0	1.00	FALSE	3	Var: 11817, Muj: 9368, No : 25, Ind: 0
Circunstancia de Contacto	0	1.00	FALSE	2	Urg: 18989, Pro: 2221, Otr: 0
Tipo Alta	324	0.98	FALSE	6	Dom: 19425, Alt: 524, Tra: 509, Tra: 368
Categoría	0	1.00	FALSE	7	Esq: 9126, Tra: 5224, Tra: 3248, Tra: 2082
Nivel Severidad APR	0	1.00	FALSE	4	1: 10666, 2: 9869, 3: 526, 4: 149
Riesgo Mortalidad APR	0	1.00	FALSE	4	1: 20197, 2: 854, 3: 122, 4: 37
Servicio	0	1.00	FALSE	30	PSQ: 19798, MIR: 547, NRL: 306, PED: 219
Régimen Financiación	21210	0.00	FALSE	0	Seg: 0, Cor: 0, Mut: 0, Acc: 0
Tipo GRD APR	0	1.00	FALSE	2	M: 21081, Q: 129
Mes de Ingreso	0	1.00	FALSE	36	201: 689, 201: 684, 201: 677, 201: 664

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Estancia Días	0	1	15.46	19.88	0.00	5.00	11.00	19.00	814.00	
GRD APR	0	1	751.34	33.58	4.00	750.00	752.00	753.00	952.00	
CDM APR	0	1	18.99	0.95	0.00	19.00	19.00	19.00	24.00	
Edad	0	1	43.64	14.11	0.00	34.00	44.00	53.00	96.00	
Coste APR	0	1	5453.11	1561.75	1496.00	4228.00	5988.00	6319.00	70601.00	
CIE	0	1	10.00	0.00	10.00	10.00	10.00	10.00	10.00	
Peso Español APR	0	1	1.20	0.34	0.33	0.93	1.32	1.39	15.52	
Edad en Ingreso	0	1	43.68	14.12	0.00	34.00	44.00	53.00	96.00	

Como se puede observar en el resumen acerca de los valores nulos, hay muchas columnas con campos vacíos. Para verlo con más facilidad, se hará el porcentaje de valores nulos para cada columna.

```
# Porcentaje de NA en cada columna
porcentaje_na_columnas <- sapply(SaludMental, function(x) round(sum(is.na(x)) / length(x) * 100, 2))
head(porcentaje_na_columnas,20)
```

```
##        Comunidad Autónoma                    Nombre       Fecha de nacimiento
##                      0.00                      0.00                      0.00
##                      Sexo           CCAA Residencia         Fecha de Ingreso
##                      0.00                    100.00                      0.00
## Circunstancia de Contacto     Fecha de Fin Contacto                 Tipo Alta
```

```
##                     0.00                     0.00                     1.53
##            Estancia Días    Diagnóstico Principal                Categoría
##                     0.00                     0.00                     0.00
##            Diagnóstico 2            Diagnóstico 3            Diagnóstico 4
##                    12.28                    29.00                    45.87
##            Diagnóstico 5            Diagnóstico 6            Diagnóstico 7
##                    60.65                    72.83                    81.92
##            Diagnóstico 8            Diagnóstico 9
##                    88.10                    92.38
```

De esta manera también se puede observar que hay varias columnas con un alto porcentaje de valores nulos, como pueden ser muchos de los diagnósticos, además de varias columnas con todos los valores nulos, como *CCAA Residencia* o un gran número de los procedimientos. Para evitar esto, existen varias técnicas de imputación de datos que emplea, como la sustitución por la media o la moda, por el mínimo valor o eliminar el registro con el valor nulo.

Tratamiento de valores nulos

En primer lugar, se hace un primer diagnóstico acerca de los posibles valores a imputar

```
diagnose(SaludMental)
```

```
## # A tibble: 111 x 6
##    variables        types missing_count missing_percent unique_count unique_rate
##    <chr>            <chr>         <int>           <dbl>        <int>       <dbl>
##  1 Comunidad Autón~ fact~             0               0            2   0.0000943
##  2 Nombre           char~             0               0        12455   0.587
##  3 Fecha de nacimi~ POSI~             0               0         9302   0.439
##  4 Sexo             fact~             0               0            3   0.000141
##  5 CCAA Residencia  logi~         21210             100            1   0.0000471
##  6 Fecha de Ingreso POSI~             0               0         1096   0.0517
##  7 Circunstancia d~ fact~             0               0            2   0.0000943
##  8 Fecha de Fin Co~ char~             0               0         1133   0.0534
##  9 Tipo Alta        fact~           324            1.53            7   0.000330
## 10 Estancia Días    nume~             0               0          175   0.00825
## # i 101 more rows
```

Como se ha mencionado, existen varias columnas sin datos, por los que se borrarán algunas de ellas, así como otras columnas redundantes.

```
SaludMental |>
  select(
    where(
      ~ all(is.na(.)) # Condición: donde TODOS los valores (all) son NA en esa columna (.)
    )
  ) |>
  names()
```

```
##  [1] "CCAA Residencia"          "Procedimiento 12"
##  [3] "Procedimiento 13"         "Procedimiento 14"
##  [5] "Procedimiento 15"         "Procedimiento 16"
```

```
##  [7] "Procedimiento 17"         "Procedimiento 18"
##  [9] "Procedimiento 19"         "Procedimiento 20"
## [11] "GDR AP"                   "CDM AP"
## [13] "Tipo GDR AP"              "Valor Peso Español"
## [15] "Tipo GDR APR"             "Valor Peso Americano APR"
## [17] "Reingreso"                "GDR IR"
## [19] "Tipo GDR IR"              "Tipo PROCESO IR"
## [21] "Régimen Financiación"     "Procedimiento Externo 1"
## [23] "Procedimiento Externo 2"  "Procedimiento Externo 3"
## [25] "Procedimiento Externo 4"  "Procedimiento Externo 5"
## [27] "Procedimiento Externo 6"
```

```r
# nulas
SaludMental$`CCAA Residencia` <- NULL
SaludMental$`Valor Peso Español`<- NULL
SaludMental$`Valor Peso Americano APR`<- NULL
SaludMental$`Tipo PROCESO IR`<- NULL

# redundantes
SaludMental$`Mes de Ingreso`<- NULL
SaludMental$`Edad en Ingreso`<- NULL
SaludMental$Procedencia <- NULL
```

Por otro lado procederemos a imputar valores nulos en columnas relevantes

```r
SaludMental$Sexo[is.na(SaludMental$Sexo)]  <- "No especificado"
```

Otras columnas significativas son datos médicos o acerca de diagnósticos y procedimientos, por los que para esos no se realizarán imputaciones.

Por otro lado, eliminaremos los registros para los que no existe *CIP SNS Recodificado*, el cual es un código identificación personal.

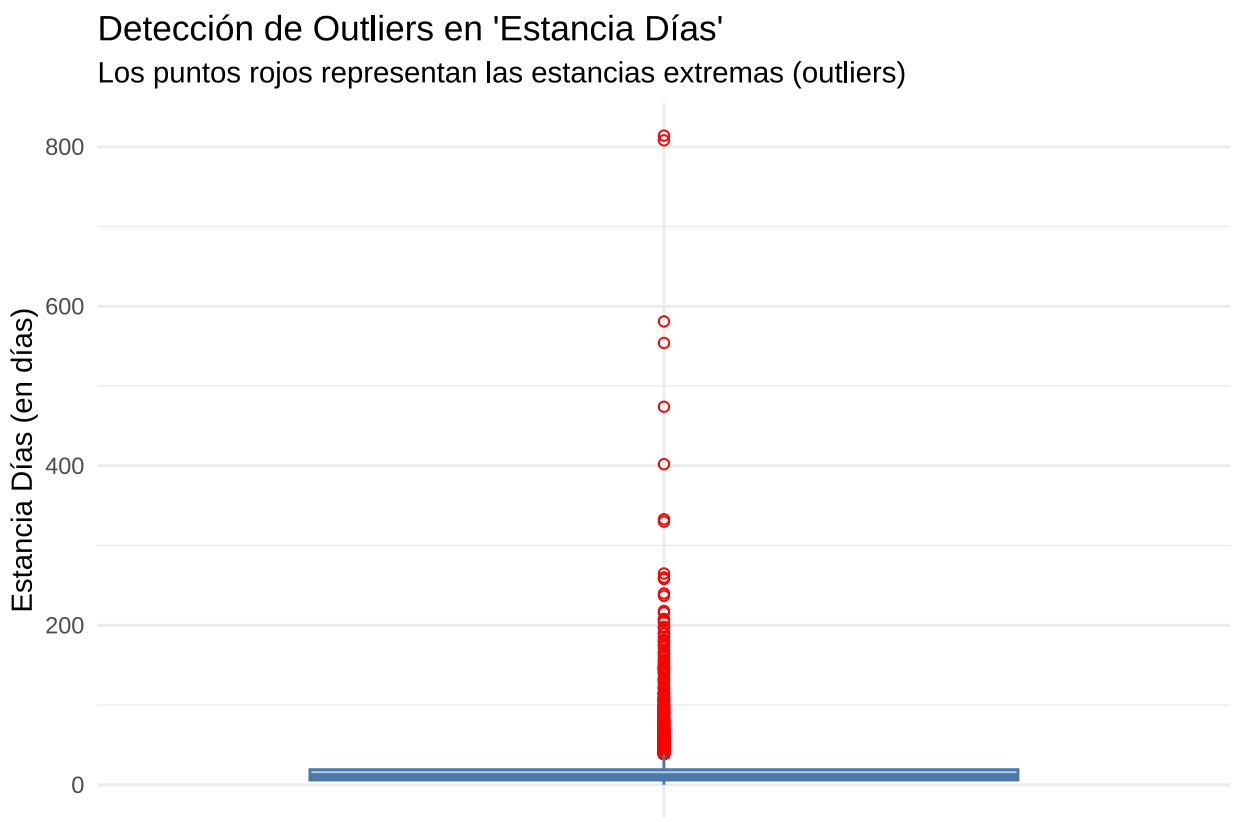
```r
SaludMental <- SaludMental[!is.na(SaludMental$`CIP SNS Recodificado`),]
```

Detección y Análisis de Outliers

Como se ha mencionado, existen variables con outliers potenciales. Para visualizarlos mejor, vamos a prepresentarlos con un diagrama de caja y bigotes en el caso de las variables numéricas y un histograma en el caso de las categóricas.

```r
SaludMental |>
  ggplot(aes(y = `Estancia Días`, x = "")) +
  geom_boxplot(
    fill = "#A8C6E1",      # Color de relleno de la caja
    color = "#4C78A8",     # Color del borde
    outlier.color = "red", # Colorear los outliers en rojo
    outlier.shape = 1      # Dar una forma específica (círculo hueco)
  ) +
  labs(
    title = "Detección de Outliers en 'Estancia Días'",
    subtitle = "Los puntos rojos representan las estancias extremas (outliers)",
```

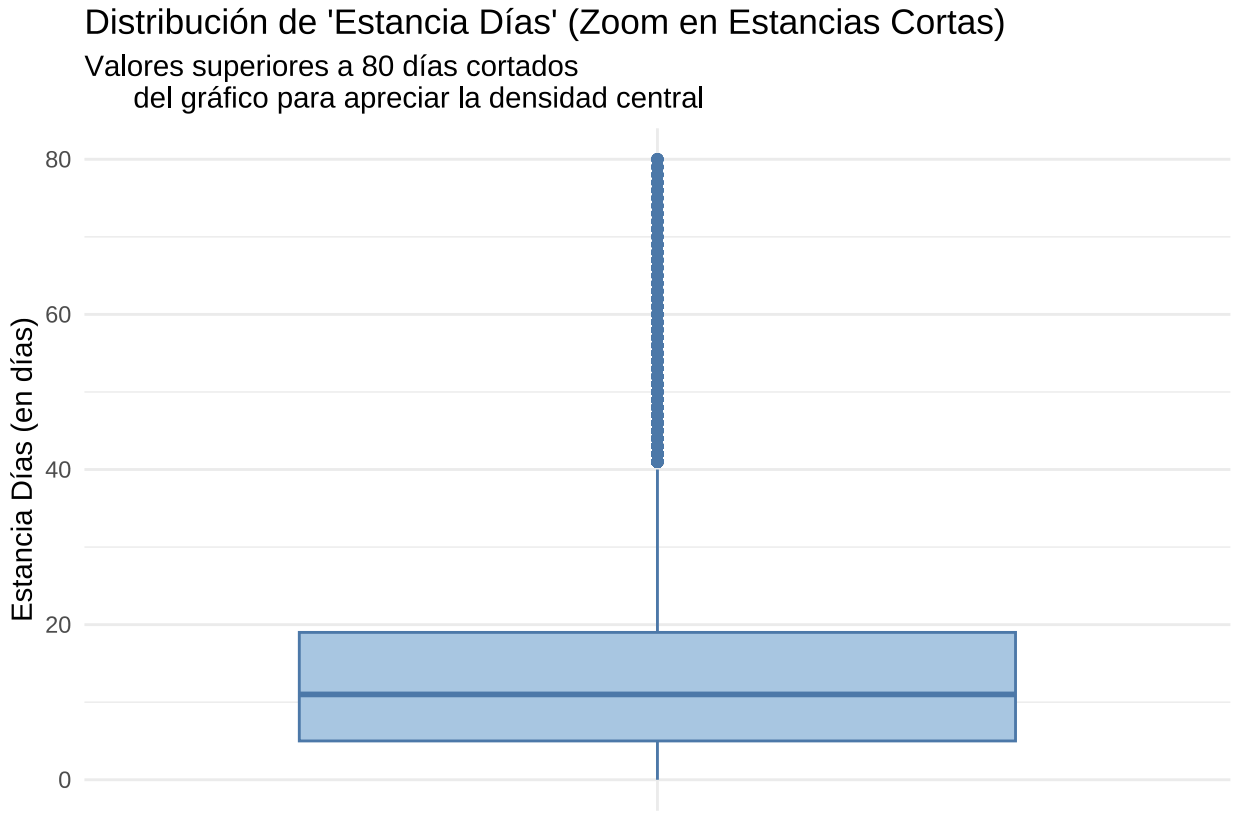
```
    y = "Estancia Días (en días)",
    x = ""
  ) +
  # Eliminar etiquetas redundantes del eje X
  theme_minimal() +
  theme(axis.text.x = element_blank())
```



Debido a que los outliers mencionados están muy alejados de la media no se aprecia bien el diagrama de caja y bigotes, por lo que se mostrará una sección de este.

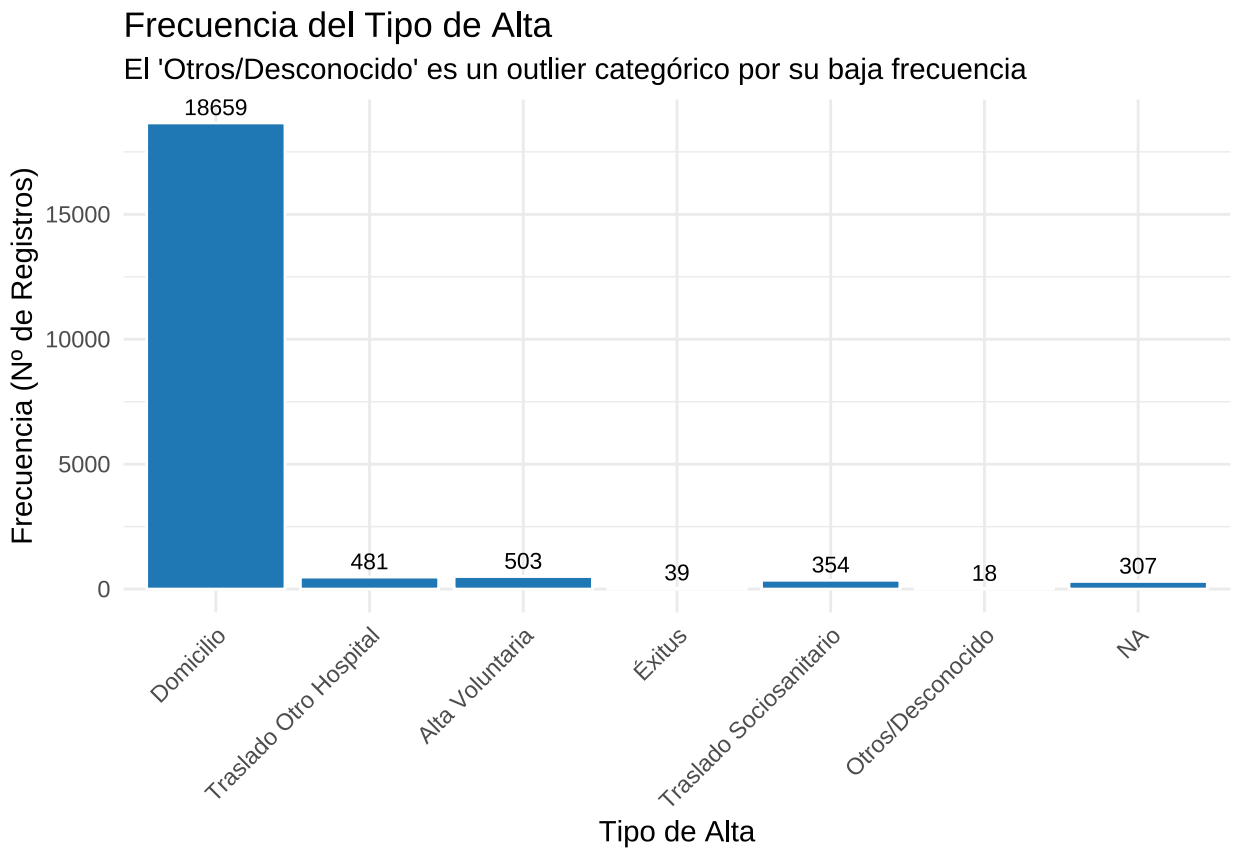
```
SaludMental |>
  ggplot(aes(y = `Estancia Días`, x = "")) +
  geom_boxplot(fill = "#A8C6E1", color = "#4C78A8") +

  # Añadimos un límite en el eje Y para hacer "zoom" en el cuerpo de la distribución
  # Por ejemplo, para ver hasta el percentil 95 (que suele estar alrededor de 40-50 días)
  ylim(0, 80) +

  labs(
    title = "Distribución de 'Estancia Días' (Zoom en Estancias Cortas)",
    subtitle = "Valores superiores a 80 días cortados
      del gráfico para apreciar la densidad central",
    y = "Estancia Días (en días)",
    x = ""
  ) +
```

```
  theme_minimal() +
  theme(axis.text.x = element_blank())
```



De la misma forma, se hará un diagrama que refleje outliers en una variable categórica.

```
SaludMental |>
  # Usamos la variable Tipo Alta ya etiquetada como factor
  ggplot(aes(x = `Tipo Alta`)) +
  geom_bar(fill = "#1F77B4", color = "white") +
  # Añadimos las etiquetas de conteo encima de las barras
  geom_text(stat = "count", aes(label = after_stat(count)), vjust = -0.5, size = 3) +
  labs(
    title = "Frecuencia del Tipo de Alta",
    subtitle = "El 'Otros/Desconocido' es un outlier categórico por su baja frecuencia",
    x = "Tipo de Alta",
    y = "Frecuencia (Nº de Registros)"
  ) +
  theme_minimal() +
  # Rotar texto del eje X para mejor lectura
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Ingeniería de características

Variables Temporales y Demográficas

En este apartado vamos a agrupar las muestras de variables tanto demográficas como temporales.

Empezaremos segmentando la variable *Edad* en 3 grupos: Menores de Edad, Adultos hasta 45, Adultos hasta 65 y Mayores

```
SaludMental <-  SaludMental |>
  mutate(Edad_Rango =
           case_when( SaludMental$Edad < 18 ~ "Menor (0-17)",
                                SaludMental$Edad >= 18 & SaludMental$Edad < 45 ~ "Adulto (18-44)",
                                SaludMental$Edad >= 45 & SaludMental$Edad < 65 ~ "Adulto (45-64)",
                      TRUE ~ "Mayor (65+)" ))


SaludMental |>
  select(Nombre, `CIP SNS Recodificado`, Edad, Edad_Rango)
```

```
## # A tibble: 20,361 x 4
##    Nombre                       `CIP SNS Recodificado`  Edad Edad_Rango
##    <chr>                        <chr>                  <dbl> <chr>
##  1 MONICA TINEO RODRIGUEZ       109457269-593755146       64 Adulto (45-64)
```

```
##  2 IRENE RODRIGUEZ HERNANDEZ     -1589750168781380096      86 Mayor (65+)
##  3 JOSE MORILLO GONZALEZ         -5406560181117020160      39 Adulto (18-44)
##  4 ELIZABETH MARTIN GUTIERREZ    -1823171082               39 Adulto (18-44)
##  5 MARIA ENCARNACION VEGA GARCIA -2828047377               38 Adulto (18-44)
##  6 ANTONIO BAUTISTA NAVARRO      -1085351946729890048      29 Adulto (18-44)
##  7 ANA ISABEL CABRERA CONTRERAS  761207796-2071749362      20 Adulto (18-44)
##  8 NEREA VAZQUEZ RODRIGUEZ       1833036024629461134       51 Adulto (45-64)
##  9 ALVARO ROSA TORRES            -1568575502649100032      49 Adulto (45-64)
## 10 REMEDIOS HUERTAS JIMENEZ      1197274748-946628087      28 Adulto (18-44)
## # i 20,351 more rows
```

Por otro lado, también puede ser interesante añadir el día de la semana con el propósito de evaluar si los los ingresos varían según el día.

```
SaludMental <- SaludMental |>
  mutate(
      Dia_Semana_Ingreso = wday(`Fecha de Ingreso`, label = TRUE, abbr = FALSE),
      Dia_Semana_Ingreso = as.factor(Dia_Semana_Ingreso)
    )


SaludMental |>
  select(Nombre, `CIP SNS Recodificado`, `Fecha de Ingreso`, Dia_Semana_Ingreso)
```

```
## # A tibble: 20,361 x 4
##    Nombre          `CIP SNS Recodificado` `Fecha de Ingreso`  Dia_Semana_Ingreso
##    <chr>           <chr>                  <dttm>              <ord>
##  1 MONICA TINEO R~ 109457269-593755146    2016-01-01 00:00:00 viernes
##  2 IRENE RODRIGUE~ -1589750168781380096   2016-01-01 00:00:00 viernes
##  3 JOSE MORILLO G~ -5406560181117020160   2016-01-01 00:00:00 viernes
##  4 ELIZABETH MART~ -1823171082            2016-01-01 00:00:00 viernes
##  5 MARIA ENCARNAC~ -2828047377            2016-01-01 00:00:00 viernes
##  6 ANTONIO BAUTIS~ -1085351946729890048   2016-01-01 00:00:00 viernes
##  7 ANA ISABEL CAB~ 761207796-2071749362   2016-01-01 00:00:00 viernes
##  8 NEREA VAZQUEZ ~ 1833036024629461134    2016-01-01 00:00:00 viernes
##  9 ALVARO ROSA TO~ -1568575502649100032   2016-01-01 00:00:00 viernes
## 10 REMEDIOS HUERT~ 1197274748-946628087   2016-01-01 00:00:00 viernes
## # i 20,351 more rows
```

Variables Clínicas y de Calidad

Estas variables se centran en la complejidad clínica y la calidad asistencial, utilizando la información de diagnósticos y procedimientos. En este caso, contaremos a cuántos procedimientos se ha sometido cada paciente, así como el número de diagnósticos.

```
SaludMental <- SaludMental |>

  rowwise() |> # Cambiar el modo de procesamiento a 'Fila por Fila'
  mutate(
    # Num_Diagnosticos: Cuenta valores NO nulos en las columnas que empiezan por "Diagnóstico"
    Num_Diagnosticos = sum(!is.na(c_across(starts_with("Diagnóstico")))),
```

```
    # Num_Procedimientos: Cuenta valores NO nulos en las columnas que empiezan por "Procedimiento"
    Num_Procedimientos = sum(!is.na(c_across(starts_with("Procedimiento"))))
  ) |>

  ungroup()



SaludMental |>
  select(Nombre, `CIP SNS Recodificado`, Num_Diagnosticos, Num_Procedimientos)
```

```
## # A tibble: 20,361 x 4
##    Nombre             `CIP SNS Recodificado` Num_Diagnosticos Num_Procedimientos
##    <chr>              <chr>                             <int>              <int>
##  1 MONICA TINEO RODR~ 109457269-593755146                   3                  0
##  2 IRENE RODRIGUEZ H~ -1589750168781380096                  6                  3
##  3 JOSE MORILLO GONZ~ -5406560181117020160                  2                  0
##  4 ELIZABETH MARTIN ~ -1823171082                           6                  0
##  5 MARIA ENCARNACION~ -2828047377                           2                  0
##  6 ANTONIO BAUTISTA ~ -1085351946729890048                  2                  0
##  7 ANA ISABEL CABRER~ 761207796-2071749362                  4                  0
##  8 NEREA VAZQUEZ ROD~ 1833036024629461134                   6                  0
##  9 ALVARO ROSA TORRES -1568575502649100032                  4                  0
## 10 REMEDIOS HUERTAS ~ 1197274748-946628087                  5                  0
## # i 20,351 more rows
```

Debido a que muchos de estos datos son información sensible y personal del paciente como puede ser el nombre o la fecha de nacimiento , hay que anonimizarlos para después tratarlos.

Conclusión

El análisis exploratorio realizado sobre el conjunto de datos de salud mental revela aspectos cruciales tanto en la calidad de los datos como en las características demográficas y asistenciales de los pacientes.

Calidad y Estructura de los Datos

Integridad de datos. Se identificó un porcentaje alto de valores nulos (NA) en numerosas variables, especialmente en aquellas procedimentales y de diagnóstico, lo que representa un desafío de calidad de datos. Adicionalmente, variables clave como *Reingreso* o *CCAA Residencia* muestran una gran falta de información, por lo que esas columnas se han eliminado, así como registros en los que faltaba el código identificativo. Por otro lado, se ha realizado alguna imputación de valores nulos.

Preparación de Variables. Las principales variables categóricas (*Sexo*, *Tipo Alta*, *Circunstancia de Contacto*, etc) fueron etiquetadas de acuerdo al documento porporcionado y transformadas a factores para facilitar su posterior procesamiento.

Resultados Estadísticos Clave

Distribuciones de los datos. Se han encontrado tanto variables con datos que siguen una distribución gausiana, como puede ser *Edad*, como variables con asímetrias ligeras o significativas (*Estancia Días*) provocadas por un número de outliers que distorsionan valores clave como la media.

Ingeniería de Características

Preparación para Modelado. Se ha enriquecido el dataset mediante *Feature Engineering*, creando variables esenciales como el rango de edad o recuentos de diagnóstico, con el propósito de potenciar

En resumen, el dataset está listo para la fase de modelado tras una estrategia rigurosa de imputación o limpieza de datos nulos y una consideración especial de los outliers para garantizar la validez de cualquier análisis inferencial.