

# 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>      <dtm>                <dbl> <lg1>
## 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` <dtm>,
## #   `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

summary(SaludMental)

```
## 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      Diagnóstico Principal      Categoría      Diagnóstico 2
## Min.      : 0.00      Length:21210      Length:21210      Length:21210
## 1st Qu.: 5.00      Class :character      Class :character      Class :character
## Median :11.00      Mode  :character      Mode  :character      Mode  :character
## Mean    :15.46
## 3rd Qu.:19.00
## Max.    :814.00
##
## Diagnóstico 3      Diagnóstico 4      Diagnóstico 5      Diagnóstico 6
## Length:21210      Length:21210      Length:21210      Length:21210
## Class :character      Class :character      Class :character      Class :character
## Mode  :character      Mode  :character      Mode  :character      Mode  :character
##
##
##
##
## Diagnóstico 7      Diagnóstico 8      Diagnóstico 9      Diagnóstico 10
## Length:21210      Length:21210      Length:21210      Length:21210
## Class :character      Class :character      Class :character      Class :character
## Mode  :character      Mode  :character      Mode  :character      Mode  :character
##
##
##
##
## Diagnóstico 11      Diagnóstico 12      Diagnóstico 13      Diagnóstico 14
## Length:21210      Length:21210      Length:21210      Length:21210
## Class :character      Class :character      Class :character      Class :character
## Mode  :character      Mode  :character      Mode  :character      Mode  :character
##
##
##
```

```

## Fecha de Intervención Procedimiento 1    Procedimiento 2    Procedimiento 3
## Length:21210          Length:21210      Length:21210      Length:21210
## Class :character      Class :character  Class :character  Class :character
## Mode :character       Mode :character  Mode :character   Mode :character
##
##
##
## Procedimiento 4    Procedimiento 5    Procedimiento 6    Procedimiento 7
## Length:21210      Length:21210      Length:21210      Length:21210
## Class :character   Class :character   Class :character   Class :character
## Mode :character    Mode :character    Mode :character    Mode :character
##
##
##
## Procedimiento 8    Procedimiento 9    Procedimiento 10   Procedimiento 11
## Length:21210      Length:21210      Length:21210      Length:21210
## Class :character   Class :character   Class :character   Class :character
## Mode :character    Mode :character    Mode :character    Mode :character
##
##
##
## Procedimiento 12   Procedimiento 13   Procedimiento 14   Procedimiento 15
## Mode:logical       Mode:logical       Mode:logical       Mode:logical
## NA's:21210         NA's:21210        NA's:21210        NA's:21210
##
##
##
## Procedimiento 16   Procedimiento 17   Procedimiento 18   Procedimiento 19
## Mode:logical       Mode:logical       Mode:logical       Mode:logical
## NA's:21210         NA's:21210        NA's:21210        NA's:21210
##
##
##
## Procedimiento 20   GDR AP           CDM AP           Tipo GDR AP
## Mode:logical       Mode:logical      Mode:logical      Mode:logical
## NA's:21210         NA's:21210        NA's:21210        NA's:21210
##
##
##
## Valor Peso Español   GRD APR          CDM APR          Tipo GDR APR
## Mode:logical         Min.   : 4.0      Min.   : 0.00     Mode:logical
## NA's:21210           1st Qu.:750.0    1st Qu.:19.00     NA's:21210
##                      Median :752.0    Median :19.00
##                      Mean   :751.3    Mean   :18.99
##                      3rd Qu.:753.0    3rd Qu.:19.00
##                      Max.   :952.0    Max.   :24.00
##
## Valor Peso Americano APR Nivel Severidad APR Riesgo Mortalidad APR
## Mode:logical         Min.   :1.000      Min.   :1.000
## NA's:21210           1st Qu.:1.000      1st Qu.:1.000
##                      Median :1.000      Median :1.000
##                      Mean   :1.536      Mean   :1.057

```

```

##          3rd Qu.:2.000      3rd Qu.:1.000
##          Max.    :4.000      Max.    :4.000
## Servicio      Edad      Reingreso      Coste APR
## Length:21210   Min.    : 0.00   Mode:logical   Min.    : 1496
## Class :character 1st Qu.:34.00   NA's:21210     1st Qu.: 4228
## Mode  :character Median :44.00           Median : 5988
##              Mean  :43.64           Mean  : 5453
##              3rd Qu.:53.00           3rd Qu.: 6319
##              Max.   :96.00           Max.   :70601
## GDR IR      Tipo GDR IR  Tipo PROCESO IR  CIE
## Mode:logical Mode:logical Mode:logical   Min.    :10
## NA's:21210   NA's:21210   NA's:21210     1st Qu.:10
##              Median :10
##              Mean   :10
##              3rd Qu.:10
##              Max.   :10
## Número de registro anual Centro Recodificado CIP SNS Recodificado
## Length:21210      Length:21210      Length:21210
## Class :character   Class :character   Class :character
## Mode  :character   Mode  :character   Mode  :character
##
##
## País Nacimiento  País Residencia  Fecha de Inicio contacto
## Length:21210     Length:21210     Length:21210
## Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character
##
##
## Régimen Financiación Procedencia  Continuidad Asistencial
## Length:21210      Length:21210      Length:21210
## Class :character   Class :character   Class :character
## Mode  :character   Mode  :character   Mode  :character
##
##
## Ingreso en UCI      Días UCI      Diagnóstico 15      Diagnóstico 16
## Length:21210        Length:21210      Length:21210      Length:21210
## Class :character     Class :character   Class :character   Class :character
## Mode  :character     Mode  :character   Mode  :character   Mode  :character
##
##
## Diagnóstico 17      Diagnóstico 18      Diagnóstico 19      Diagnóstico 20
## Length:21210        Length:21210      Length:21210      Length:21210
## Class :character     Class :character   Class :character   Class :character
## Mode  :character     Mode  :character   Mode  :character   Mode  :character
##
##
## POA Diagnóstico Principal POA Diagnóstico 2 POA Diagnóstico 3
## Length:21210          Length:21210      Length:21210
## Class :character       Class :character   Class :character

```

```

## Mode :character          Mode :character  Mode :character
##
##
##
## POA Diagnóstico 4 POA Diagnóstico 5 POA Diagnóstico 6 POA Diagnóstico 7
## Length:21210      Length:21210      Length:21210      Length:21210
## Class :character  Class :character  Class :character  Class :character
## Mode :character   Mode :character   Mode :character   Mode :character
##
##
##
## POA Diagnóstico 8 POA Diagnóstico 9 POA Diagnóstico 10 POA Diagnóstico 11
## Length:21210      Length:21210      Length:21210      Length:21210
## Class :character  Class :character  Class :character  Class :character
## Mode :character   Mode :character   Mode :character   Mode :character
##
##
##
## POA Diagnóstico 12 POA Diagnóstico 13 POA Diagnóstico 14 POA Diagnóstico 15
## Length:21210      Length:21210      Length:21210      Length:21210
## Class :character  Class :character  Class :character  Class :character
## Mode :character   Mode :character   Mode :character   Mode :character
##
##
##
## POA Diagnóstico 16 POA Diagnóstico 17 POA Diagnóstico 18 POA Diagnóstico 19
## Length:21210      Length:21210      Length:21210      Length:21210
## Class :character  Class :character  Class :character  Class :character
## Mode :character   Mode :character   Mode :character   Mode :character
##
##
##
## POA Diagnóstico 20 Procedimiento Externo 1 Procedimiento Externo 2
## Length:21210      Mode:logical      Mode:logical
## Class :character  NA's:21210      NA's:21210
## Mode :character
##
##
##
## Procedimiento Externo 3 Procedimiento Externo 4 Procedimiento Externo 5
## Mode:logical      Mode:logical      Mode:logical
## NA's:21210      NA's:21210      NA's:21210
##
##
##
## Procedimiento Externo 6 Tipo GRD APR      Peso Español APR  Edad en Ingreso
## Mode:logical      Length:21210      Min. : 0.3298     Min. : 0.00
## NA's:21210      Class :character  1st Qu.: 0.9255    1st Qu.:34.00
##                  Mode :character  Median : 1.3163     Median :44.00
##                  Mean : 1.1968     Mean :43.68
##                  3rd Qu.: 1.3930     3rd Qu.:53.00
##                  Max. :15.5179     Max. :96.00
## Mes de Ingreso

```

```
## Length:21210
## Class :character
## Mode :character
##
##
##
```

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
```

```
# 1. Calcular Media 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)

# 2. Generar el gráfico con la curva de Densidad (naranja) Y la Gaussiana (roja)
SaludMental |>
  ggplot(aes(x = Edad)) +
  # 1. Histograma (Muestra la frecuencia real de los datos)
  geom_histogram(
    aes(y = after_stat(density)),
    binwidth = 5,
    fill = "#4C78A8",
    color = "white",
    alpha = 0.5
  ) +
  # 2. Curva de Densidad

  # 3. Curva GAUSSIANA
  geom_function(
    fun = dnorm, # La función de densidad normal
    args = list(mean = media_edad, sd = sd_edad),
    color = "#98FF98", # Color rojo/vino para diferenciar
    linewidth = 1.2
  ) +

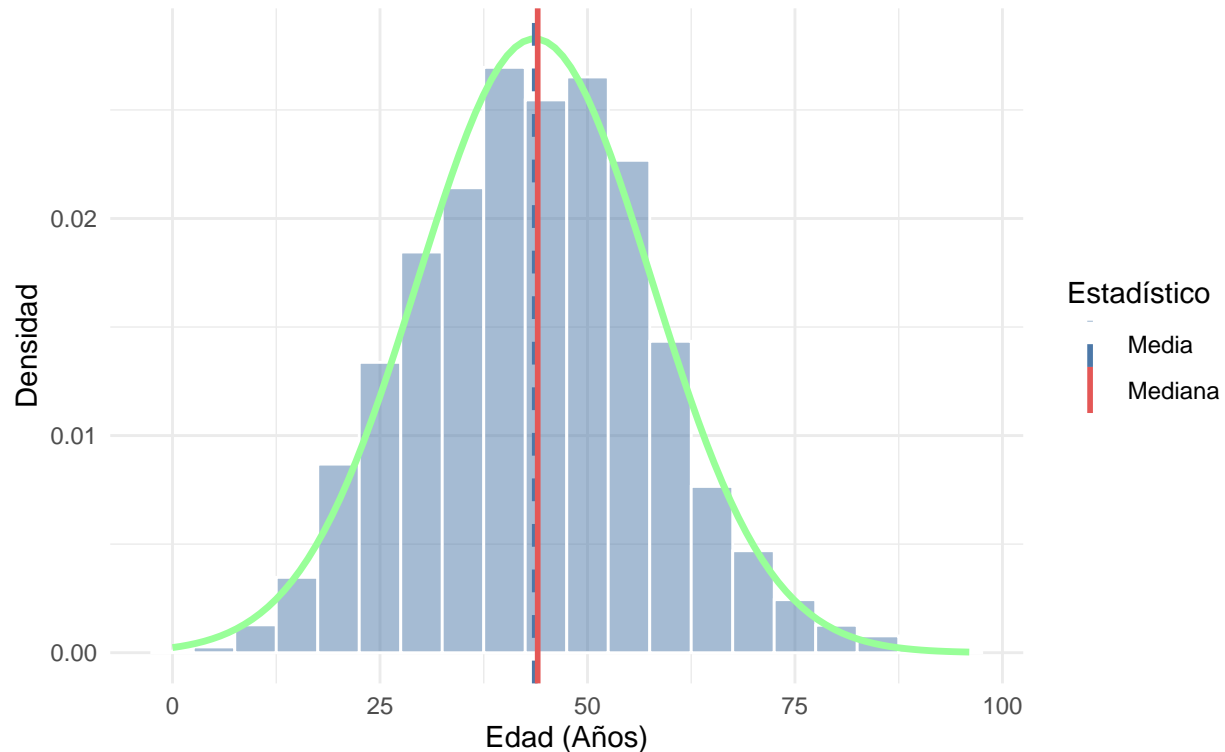
  # 4. Marcas de Media y Mediana
  geom_vline(aes(xintercept = media_edad, color = "Media"),
    linetype = "dashed", linewidth = 1) +
  geom_vline(aes(xintercept = mediana_edad, color = "Mediana"),
    linetype = "solid", linewidth = 1) +

  # 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
```

```
x = "Edad (Años)",
y = "Densidad"
) +
theme_minimal()
```

## Distribución de la Edad vs. Curva Normal Teórica

Media: 43.64 | Mediana: 44 | Desv. Estándar: 14.11



Por otro lado, en variables como *Estancia Días*, 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
```

```
# 1. 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)

# 2. Generar el gráfico de distribución
SaludMental |>
  ggplot(aes(x = `Estancia Días`)) +
  # 1. Histograma (Muestra la frecuencia real de los datos)
  geom_histogram(
    aes(y = after_stat(density)),
    binwidth = 5, # Agrupa los días en intervalos de 5
```

```

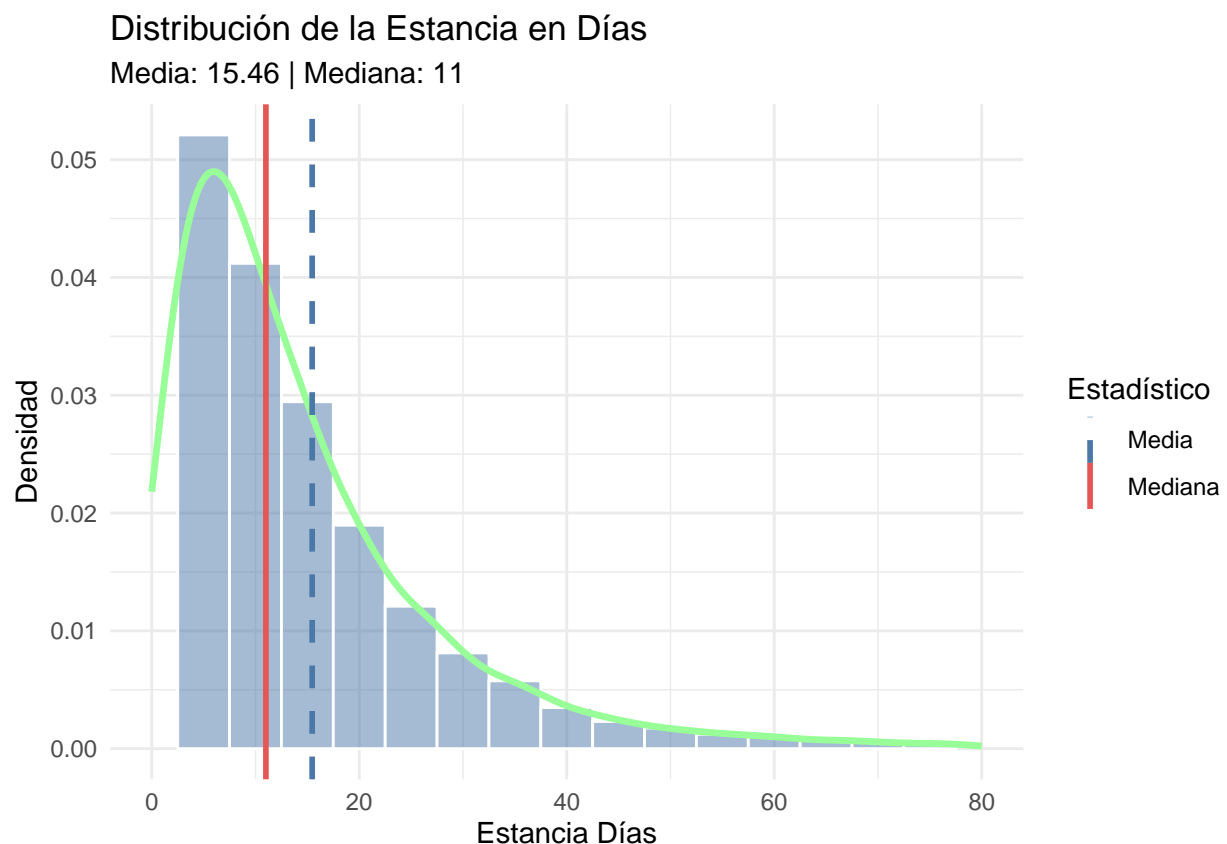
    fill = "#4C78A8",
    color = "white",
    alpha = 0.5
) +
# 2. Curva de Densidad
geom_density(linewidth = 1.2, color = "#98FF98", adjust = 2) +

# 4. 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 apreciar la asimetría y 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.

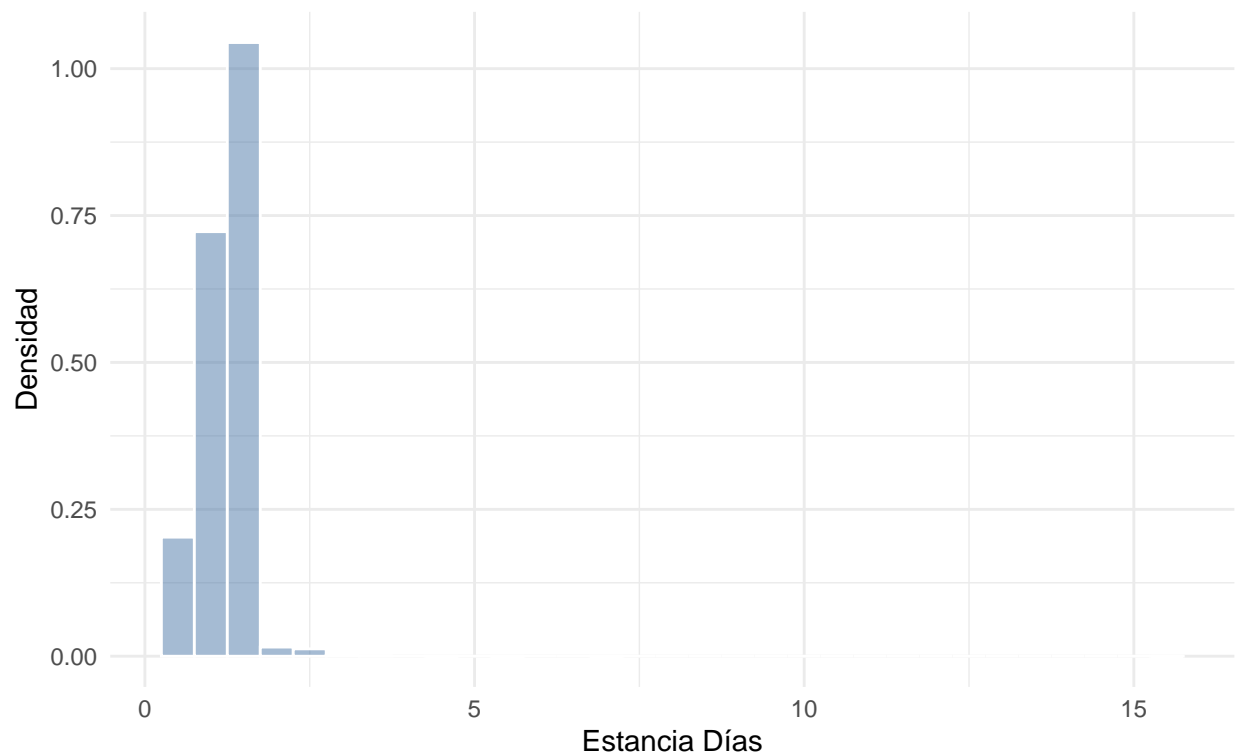
```
# 1. Calcular Media, Mediana y Desviación Estándar de Estancia Días (excluyendo NA)
media_pesoEsp <- mean(SaludMental$`Peso Español APR`, na.rm = TRUE)
mediana_pesoEsp <- median(SaludMental$`Peso Español APR`, na.rm = TRUE)

# 2. Generar el gráfico de distribución
SaludMental |>
  ggplot(aes(x = `Peso Español APR`)) +
  # 1. Histograma (Muestra la frecuencia real de los datos)
  geom_histogram(
    aes(y = after_stat(density)),
    binwidth = 0.5, # Agrupa los días en intervalos de 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 = "Estancia Días",
    y = "Densidad"
  ) +
  theme_minimal()
```

## Distribución del Peso Español APR

Media: 1.2 | Mediana: 1.32



## 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. Éxito / 5. Traslado
  `Tipo Alta` = factor(
    `Tipo Alta`,
    levels = c(1, 2, 3, 4, 5, 9),
    labels = c("Domicilio", "Traslado Otro Hospital", "Alta Voluntaria", "Éxito", "Traslado Sociosan")
  ),

  # **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", "Acciden
),

# **Circunstancia de Contacto (Asumiendo que es Tipo de Ingreso):** 1. Urgente / 2. Programado / 9.
`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.

```

# 1. Definir una lista de las variables categóricas (¡Añade todas las que sean relevantes!)
variables_categoricas <- c("Comunidad Autónoma", "Categoría", "Servicio", "Nivel Severidad APR", "Riesgo

SaludMental <- SaludMental |>
  mutate(
    across(
      .cols = all_of(variables_categoricas),
      .fns = as.factor
    )
  )
# 2. 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 ...
## $ Categoría          : Factor w/ 7 levels "Esquizofrenia, trastornos esquizotípicos y trastornos e
## $ 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

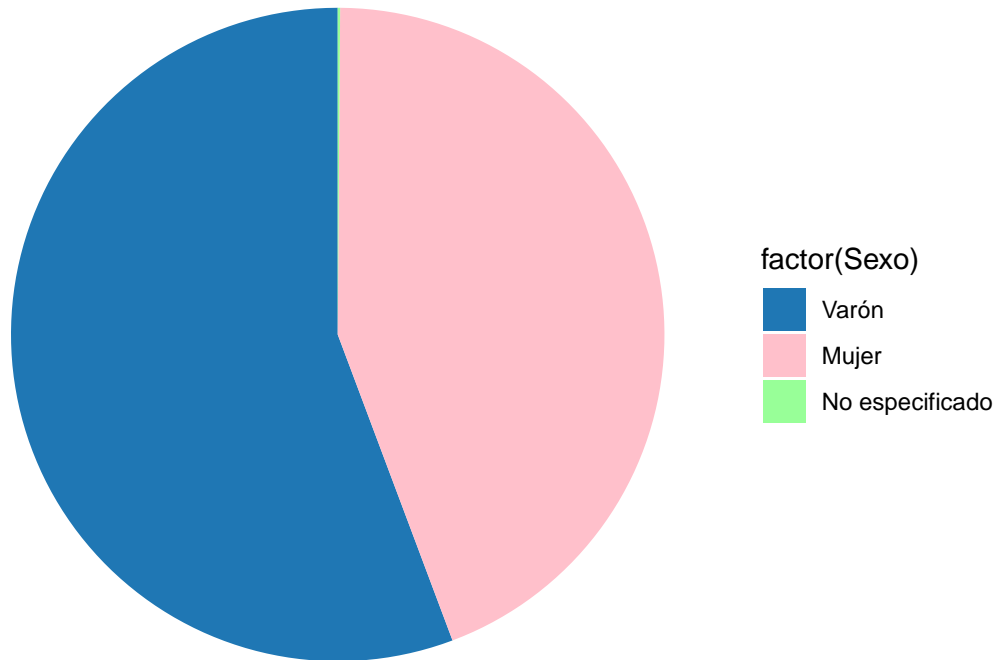
```

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) +

```

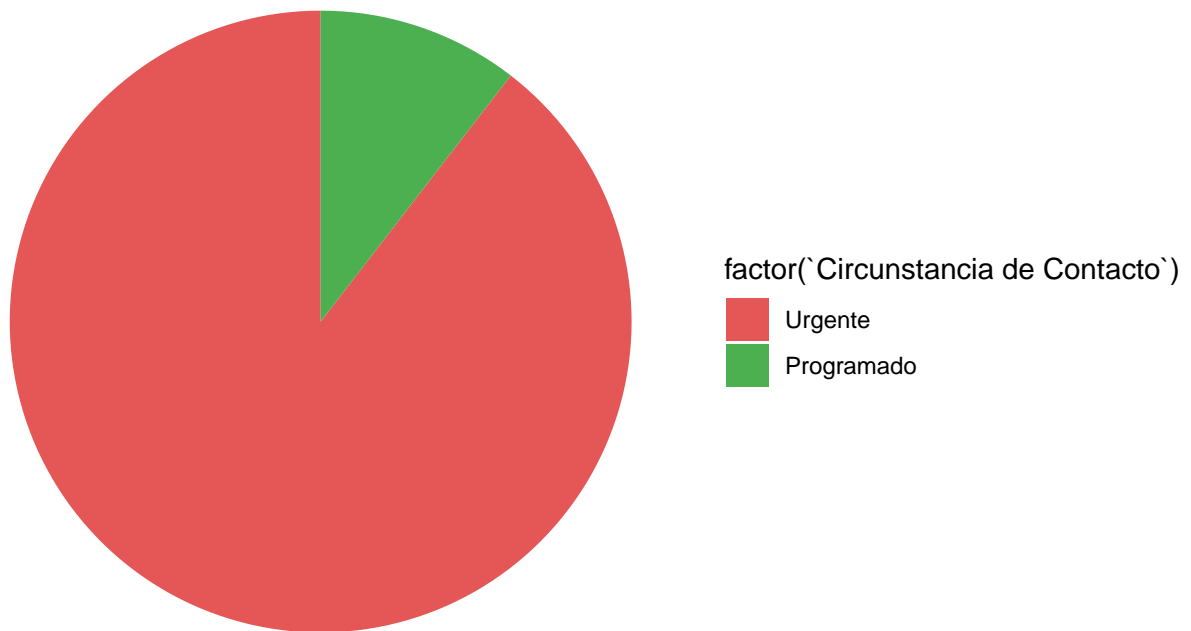
```
theme_void() +
scale_fill_manual(values = c("#1F77B4", "#FFC0CB", "#98FF98"))
```

## Distribución de Sexo



```
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"))
```

## Distribución Circunstancia de Contacto



## Estudio de Valores Desconocidos (NA)

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

```
skim(SaludMental)
```

Table 1: Data summary

Name	SaludMental
Number of rows	21210
Number of columns	111
Column type frequency:	
character	64
factor	11
logical	26
numeric	8
POSIXct	2
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Nombre	0	1.00	13	39	0	12455	0
Fecha de Fin Contacto	0	1.00	10	10	0	1133	0
Diagnóstico Principal	0	1.00	3	7	0	263	0
Diagnóstico 2	2604	0.88	3	8	0	1586	0
Diagnóstico 3	6150	0.71	3	8	0	1539	0
Diagnóstico 4	9729	0.54	3	8	0	1470	0
Diagnóstico 5	12864	0.39	3	8	0	1237	0
Diagnóstico 6	15447	0.27	3	8	0	1030	0
Diagnóstico 7	17375	0.18	3	8	0	833	0
Diagnóstico 8	18686	0.12	3	8	0	656	0
Diagnóstico 9	19593	0.08	3	8	0	518	0
Diagnóstico 10	20205	0.05	3	8	0	388	0
Diagnóstico 11	20542	0.03	3	8	0	306	0
Diagnóstico 12	20807	0.02	3	8	0	213	0
Diagnóstico 13	20969	0.01	3	8	0	132	0
Diagnóstico 14	21065	0.01	3	8	0	105	0
Fecha de Intervención	21069	0.01	13	13	0	141	0
Procedimiento 1	16590	0.22	7	7	0	365	0
Procedimiento 2	18482	0.13	7	7	0	255	0
Procedimiento 3	20136	0.05	7	7	0	177	0
Procedimiento 4	20766	0.02	7	7	0	122	0
Procedimiento 5	21017	0.01	7	7	0	74	0
Procedimiento 6	21108	0.00	7	7	0	47	0
Procedimiento 7	21140	0.00	7	7	0	33	0
Procedimiento 8	21157	0.00	7	7	0	21	0
Procedimiento 9	21173	0.00	7	7	0	15	0
Procedimiento 10	21186	0.00	7	7	0	6	0
Procedimiento 11	21194	0.00	7	7	0	7	0
Número de registro anual	0	1.00	7	10	0	21209	0
Centro Recodificado	0	1.00	10	21	0	45	0
CIP SNS Recodificado	849	0.96	9	21	0	10896	0
País Nacimiento	0	1.00	3	3	0	81	0
País Residencia	0	1.00	3	5	0	56	0
Fecha de Inicio contacto	0	1.00	13	13	0	20962	0
Procedencia	0	1.00	4	4	0	10	0
Continuidad Asistencial	0	1.00	3	3	0	7	0
Ingreso en UCI	0	1.00	3	3	0	3	0
Días UCI	21110	0.00	3	4	0	16	0
Diagnóstico 15	21119	0.00	3	8	0	70	0
Diagnóstico 16	21144	0.00	3	7	0	56	0
Diagnóstico 17	21168	0.00	3	8	0	34	0
Diagnóstico 18	21185	0.00	5	8	0	22	0
Diagnóstico 19	21191	0.00	5	8	0	16	0
Diagnóstico 20	21200	0.00	3	7	0	10	0
POA Diagnóstico Principal	0	1.00	1	1	0	4	0
POA Diagnóstico 2	2604	0.88	1	1	0	4	0
POA Diagnóstico 3	6150	0.71	1	1	0	4	0
POA Diagnóstico 4	9729	0.54	1	1	0	6	0
POA Diagnóstico 5	12864	0.39	1	1	0	5	0
POA Diagnóstico 6	15447	0.27	1	1	0	5	0
POA Diagnóstico 7	17375	0.18	1	1	0	5	0
POA Diagnóstico 8	18686	0.12	1	1	0	4	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
POA Diagnóstico 9	19593	0.08	1	1	0	4	0
POA Diagnóstico 10	20205	0.05	1	1	0	4	0
POA Diagnóstico 11	20542	0.03	1	1	0	4	0
POA Diagnóstico 12	20807	0.02	1	1	0	3	0
POA Diagnóstico 13	20969	0.01	1	1	0	3	0
POA Diagnóstico 14	21065	0.01	1	1	0	4	0
POA Diagnóstico 15	21119	0.00	1	1	0	4	0
POA Diagnóstico 16	21144	0.00	1	1	0	3	0
POA Diagnóstico 17	21168	0.00	1	1	0	3	0
POA Diagnóstico 18	21185	0.00	1	1	0	2	0
POA Diagnóstico 19	21191	0.00	1	1	0	2	0
POA Diagnóstico 20	21200	0.00	1	1	0	2	0

### 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: logical

skim_variable	n_missing	complete_rate	mean	count
CCAA Residencia	21210	0	NaN	:
Procedimiento 12	21210	0	NaN	:
Procedimiento 13	21210	0	NaN	:
Procedimiento 14	21210	0	NaN	:
Procedimiento 15	21210	0	NaN	:
Procedimiento 16	21210	0	NaN	:
Procedimiento 17	21210	0	NaN	:
Procedimiento 18	21210	0	NaN	:
Procedimiento 19	21210	0	NaN	:
Procedimiento 20	21210	0	NaN	:

skim_variable	n_missing	complete_rate	mean	count
GDR AP	21210	0	NaN	:
CDM AP	21210	0	NaN	:
Tipo GDR AP	21210	0	NaN	:
Valor Peso Español	21210	0	NaN	:
Tipo GDR APR	21210	0	NaN	:
Valor Peso Americano APR	21210	0	NaN	:
Reingreso	21210	0	NaN	:
GDR IR	21210	0	NaN	:
Tipo GDR IR	21210	0	NaN	:
Tipo PROCESO IR	21210	0	NaN	:
Procedimiento Externo 1	21210	0	NaN	:
Procedimiento Externo 2	21210	0	NaN	:
Procedimiento Externo 3	21210	0	NaN	:
Procedimiento Externo 4	21210	0	NaN	:
Procedimiento Externo 5	21210	0	NaN	:
Procedimiento Externo 6	21210	0	NaN	:

#### Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Estancia	0	1	15.46	19.88	0.00	5.00	11.00	19.00	814.00	
Días										
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	0	1	1.20	0.34	0.33	0.93	1.32	1.39	15.52	
Español										
APR										
Edad en	0	1	43.68	14.12	0.00	34.00	44.00	53.00	96.00	
Ingreso										

#### Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
Fecha de nacimiento	0	1	1921-03-07	2018-09-19	1973-03-03	9302
Fecha de Ingreso	0	1	2016-01-01	2018-12-31	2017-06-30	1096

Como se puede observar en el resumen acerca de los valores nulos, hay muchas columnas con columnas vacías. 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) sum(is.na(x)) / length(x) * 100)
porcentaje_na_columnas
```

```
## Comunidad Autónoma Nombre Fecha de nacimiento
## 0.000000 0.000000 0.000000
```



##	Sexo	CCAA Residencia	Fecha de Ingreso
##	0.000000	100.000000	0.000000
##	Circunstancia de Contacto	Fecha de Fin Contacto	Tipo Alta
##	0.000000	0.000000	1.527581
##	Estancia Días	Diagnóstico Principal	Categoría
##	0.000000	0.000000	0.000000
##	Diagnóstico 2	Diagnóstico 3	Diagnóstico 4
##	12.277228	28.995757	45.869873
##	Diagnóstico 5	Diagnóstico 6	Diagnóstico 7
##	60.650636	72.828854	81.918906
##	Diagnóstico 8	Diagnóstico 9	Diagnóstico 10
##	88.099953	92.376238	95.261669
##	Diagnóstico 11	Diagnóstico 12	Diagnóstico 13
##	96.850542	98.099953	98.863744
##	Diagnóstico 14	Fecha de Intervención	Procedimiento 1
##	99.316360	99.335219	78.217822
##	Procedimiento 2	Procedimiento 3	Procedimiento 4
##	87.138142	94.936351	97.906648
##	Procedimiento 5	Procedimiento 6	Procedimiento 7
##	99.090052	99.519095	99.669967
##	Procedimiento 8	Procedimiento 9	Procedimiento 10
##	99.750118	99.825554	99.886846
##	Procedimiento 11	Procedimiento 12	Procedimiento 13
##	99.924564	100.000000	100.000000
##	Procedimiento 14	Procedimiento 15	Procedimiento 16
##	100.000000	100.000000	100.000000
##	Procedimiento 17	Procedimiento 18	Procedimiento 19
##	100.000000	100.000000	100.000000
##	Procedimiento 20	GDR AP	CDM AP
##	100.000000	100.000000	100.000000
##	Tipo GDR AP	Valor Peso Español	GRD APR
##	100.000000	100.000000	0.000000
##	CDM APR	Tipo GDR APR	Valor Peso Americano APR
##	0.000000	100.000000	100.000000
##	Nivel Severidad APR	Riesgo Mortalidad APR	Servicio
##	0.000000	0.000000	0.000000
##	Edad	Reingreso	Coste APR
##	0.000000	100.000000	0.000000
##	GDR IR	Tipo GDR IR	Tipo PROCESO IR
##	100.000000	100.000000	100.000000
##	CIE	Número de registro anual	Centro Recodificado
##	0.000000	0.000000	0.000000
##	CIP SNS Recodificado	País Nacimiento	País Residencia
##	4.002829	0.000000	0.000000
##	Fecha de Inicio contacto	Régimen Financiación	Procedencia
##	0.000000	100.000000	0.000000
##	Continuidad Asistencial	Ingreso en UCI	Días UCI
##	0.000000	0.000000	99.528524
##	Diagnóstico 15	Diagnóstico 16	Diagnóstico 17
##	99.570957	99.688826	99.801980
##	Diagnóstico 18	Diagnóstico 19	Diagnóstico 20
##	99.882131	99.910420	99.952852
##	POA Diagnóstico Principal	POA Diagnóstico 2	POA Diagnóstico 3
##	0.000000	12.277228	28.995757

##	POA Diagnóstico 4	POA Diagnóstico 5	POA Diagnóstico 6
##	45.869873	60.650636	72.828854
##	POA Diagnóstico 7	POA Diagnóstico 8	POA Diagnóstico 9
##	81.918906	88.099953	92.376238
##	POA Diagnóstico 10	POA Diagnóstico 11	POA Diagnóstico 12
##	95.261669	96.850542	98.099953
##	POA Diagnóstico 13	POA Diagnóstico 14	POA Diagnóstico 15
##	98.863744	99.316360	99.570957
##	POA Diagnóstico 16	POA Diagnóstico 17	POA Diagnóstico 18
##	99.688826	99.801980	99.882131
##	POA Diagnóstico 19	POA Diagnóstico 20	Procedimiento Externo 1
##	99.910420	99.952852	100.000000
##	Procedimiento Externo 2	Procedimiento Externo 3	Procedimiento Externo 4
##	100.000000	100.000000	100.000000
##	Procedimiento Externo 5	Procedimiento Externo 6	Tipo GRD APR
##	100.000000	100.000000	0.000000
##	Peso Español APR	Edad en Ingreso	Mes de Ingreso
##	0.000000	0.000000	0.000000

De esta manera podemos 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.

## Detección y Análisis de Outliers

Como se ha mencionado, existen variables con outliers potenciales. Para visualizarlos mejor, vamos a pre-resentarlos con un diagrama de caja y bigotes.

```
# Usando la librería ggplot2 que ya estás cargando
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())
```

## Detección de Outliers en 'Estancia Días'

Los puntos rojos representan las estancias extremas (outliers)



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

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

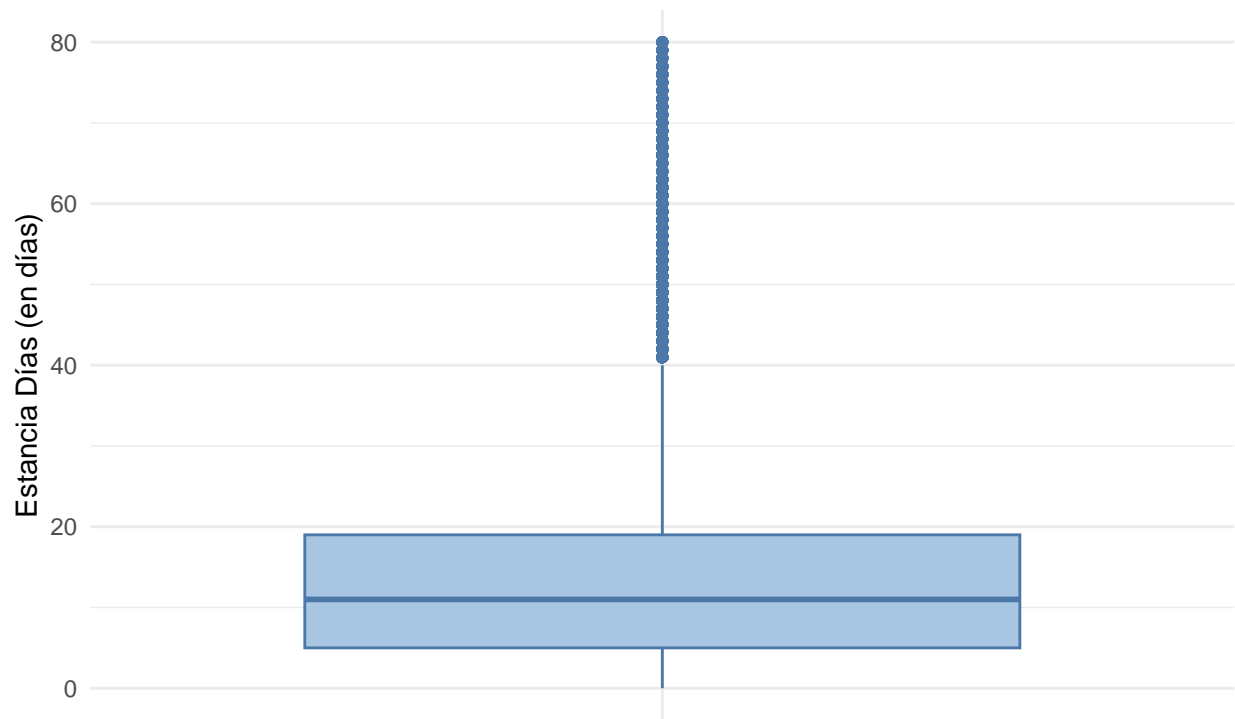
  # Añadimos un límite en el eje Y para hacer "zoom" en el cuerpo de la distribución
  # Por ejemplo, para ver hasta el percentil 95 (que suele estar alrededor de 40-50 días)
  ylim(0, 80) +

  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())
```

```
## Warning: Removed 217 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

## Distribución de 'Estancia Días' (Zoom en Estancias Cortas)

Valores superiores a 80 días cortados del gráfico para apreciar la densidad central

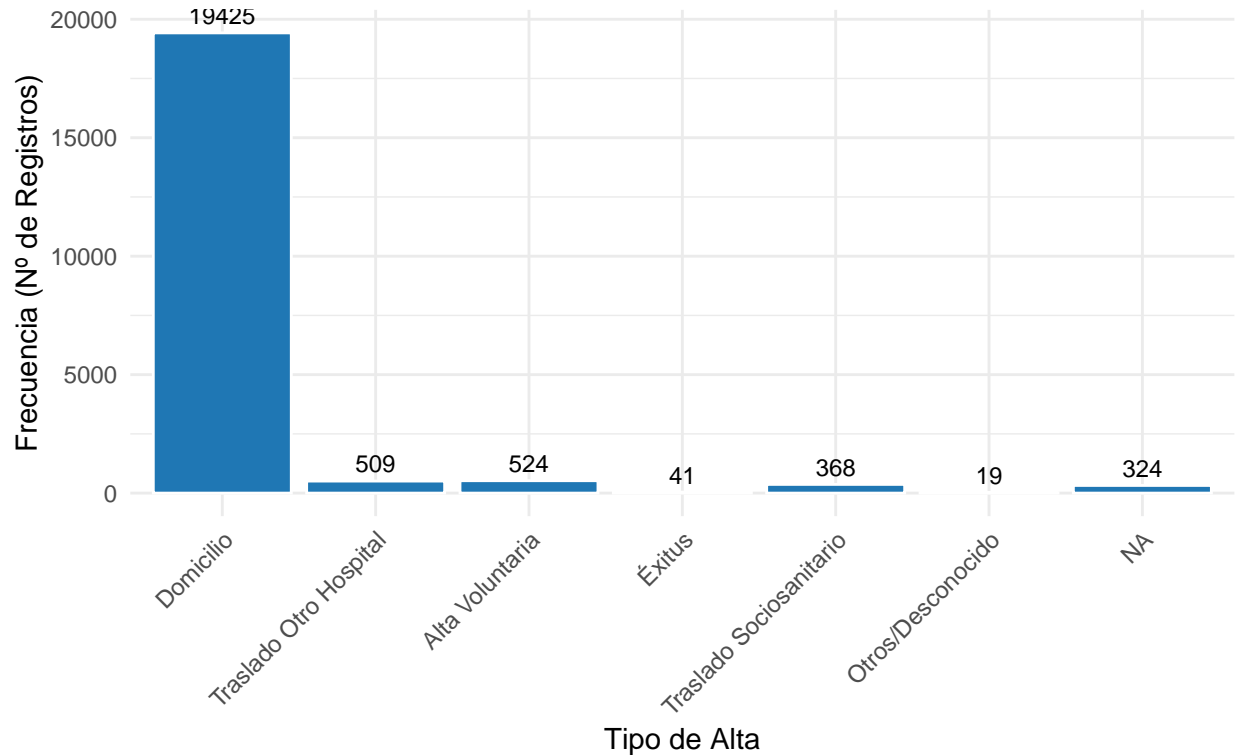


De la misma forma, se hará un diagrama que refleje outliers en una variable categórica.

```
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 '9: 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))
```

## Frecuencia del Tipo de Alta

El '9: Otros/Desconocido' es un outlier categórico por su baja frecuencia



## 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 ~ "Adulto (45-64)", SaludMental$Edad >= 65 ~ "Mayor (65+)" ))

SaludMental |>
  select(Nombre, Edad, Edad_Rango)
```

```
## # A tibble: 21,210 x 3
##   Nombre                      Edad Edad_Rango
##   <chr>                      <dbl> <chr>
## 1 MONICA TINEO RODRIGUEZ      64 Adulto (45-64)
## 2 IRENE RODRIGUEZ HERNANDEZ  86 Mayor (65+)
## 3 JOSE MORILLO GONZALEZ      39 Adulto (18-44)
## 4 ELIZABETH MARTIN GUTIERREZ 39 Adulto (18-44)
## 5 MARIA ENCARNACION VEGA GARCIA 38 Adulto (18-44)
```

```
## 6 ANTONIO BAUTISTA NAVARRO      29 Adulto (18-44)
## 7 ANA ISABEL CABRERA CONTRERAS  20 Adulto (18-44)
## 8 NEREA VAZQUEZ RODRIGUEZ      51 Adulto (45-64)
## 9 ALVARO ROSA TORRES           49 Adulto (45-64)
## 10 REMEDIOS HUERTAS JIMENEZ     28 Adulto (18-44)
## # i 21,200 more rows
```

Por otro lado, también puede ser interesante añadir el día de la semana con el propósito de evaluar si 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, `Fecha de Ingreso`, Dia_Semana_Ingreso)
```

```
## # A tibble: 21,210 x 3
##   Nombre                                `Fecha de Ingreso` Dia_Semana_Ingreso
##   <chr>                                <dtm>              <ord>
## 1 MONICA TINEO RODRIGUEZ              2016-01-01 00:00:00 viernes
## 2 IRENE RODRIGUEZ HERNANDEZ          2016-01-01 00:00:00 viernes
## 3 JOSE MORILLO GONZALEZ               2016-01-01 00:00:00 viernes
## 4 ELIZABETH MARTIN GUTIERREZ          2016-01-01 00:00:00 viernes
## 5 MARIA ENCARNACION VEGA GARCIA       2016-01-01 00:00:00 viernes
## 6 ANTONIO BAUTISTA NAVARRO            2016-01-01 00:00:00 viernes
## 7 ANA ISABEL CABRERA CONTRERAS        2016-01-01 00:00:00 viernes
## 8 NEREA VAZQUEZ RODRIGUEZ            2016-01-01 00:00:00 viernes
## 9 ALVARO ROSA TORRES                  2016-01-01 00:00:00 viernes
## 10 REMEDIOS HUERTAS JIMENEZ           2016-01-01 00:00:00 viernes
## # i 21,200 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, Num_Diagnosticos, Num_Procedimientos)
```

```
## # A tibble: 21,210 x 3  
##   Nombre                               Num_Diagnosticos Num_Procedimientos  
##   <chr>                                <int>             <int>  
## 1 MONICA TINEO RODRIGUEZ                3                 0  
## 2 IRENE RODRIGUEZ HERNANDEZ             6                 3  
## 3 JOSE MORILLO GONZALEZ                 2                 0  
## 4 ELIZABETH MARTIN GUTIERREZ            6                 0  
## 5 MARIA ENCARNACION VEGA GARCIA         2                 0  
## 6 ANTONIO BAUTISTA NAVARRO              2                 0  
## 7 ANA ISABEL CABRERA CONTRERAS         4                 0  
## 8 NEREA VAZQUEZ RODRIGUEZ              6                 0  
## 9 ALVARO ROSA TORRES                   4                 0  
## 10 REMEDIOS HUERTAS JIMENEZ             5                 0  
## # i 21,200 more rows
```