“AÑO DE LA RECUPERACIÓN Y CONSOLIDACIÓN DE LA ECONOMÍA PERUANA”

UNIVERSIDAD PRIVADA SAN JUAN BAUTISTA

FACULTAD DE CIENCIAS DE LA SALUD

ESCUELA PROFESIONAL DE MEDICINA HUMANA-EPMH

ASIGNATURA:

“SISTEMATIZACIÓN Y MÉTODOS ESTADÍSTICOS”

INTEGRANTES:

MARÍA LUCIA JACOBO ATUNCAR

GAMBOA CANALES MARIPAZ

ARIANA ABIGIAL VIDAL ROMUCHO

FERNANDA GIANELLA CASTILLA SALVADOR

SEBASTIAN PALOMINO ROJAS

KRISTY STEFANY ALVAREZ PEVES

DOCENTE:

DR. SEGUNDO VICENTE CASTRO LOPEZ

SAN BORJA

2025-2

Instalar paquetes

```{r}
# install.packages("tidyverse")
# install.packages("rio")
# install.packages("here")
# install.packages("janitor")
# install.packages("skimr")
# install.packages("visdat")
```

Cargar paquetes

```{r}
library(tidyverse)
library(rio)
library(here)
library(janitor)
```

Importando cirrosis.csv usando el paquete rio

```{r}
cirrosis= import(here("cirrosis.csv"))
```

Vistazo al contenido

¿Cuántas variables y observaciones hay?

El primer número indica el número de filas, el segundo, el número de columnas.

```{r}
dim(cirrosis)
```

```
[1] 418  20
```

¿Cuantas y qué tipos de variables hay?

```{r}
str(cirrosis)
```

```
'data.frame':   418 obs. of  20 variables:
 $ ID                : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Dias_Seguimiento  : int  400 4500 1012 1925 1504 2503 1832 2466 2400 51 ...
 $ Estado            : chr  "Fallecido" "Censurado" "Fallecido" "Fallecido" ...
 $ Medicamento       : chr  "D_penicilamina" "D_penicilamina" "D_penicilamina" "D_penicilamina" ...
 $ Edad              : int  21464 20617 25594 19994 13918 24201 20284 19379 15526 25772 ...
 $ Sexo              : chr  "Mujer" "Mujer" "Hombre" "Mujer" ...
 $ Ascitis           : chr  "Sí" "No" "No" "No" ...
 $ Hepatomegalia     : chr  "Sí" "Sí" "No" "Sí" ...
 $ Aracnoides        : chr  "Sí" "Sí" "No" "Sí" ...
 $ Edema             : chr  "Severo" "Ausente" "Leve" "Leve" ...
 $ Bilirrubina       : num  14.5 1.1 1.4 1.8 3.4 0.8 1 0.3 3.2 12.6 ...
 $ Colesterol        : int  261 302 176 244 279 248 322 280 562 200 ...
 $ Albumina          : num  2.6 4.14 3.48 2.54 3.53 3.98 4.09 4 3.08 2.74 ...
 $ Cobre             : int  156 54 210 64 143 50 52 52 79 140 ...
```

Una función similar

```{r}
dplyr::glimpse(cirrosis)
```

```
Rows: 418
Columns: 20
$ ID                <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,
16…
$ Dias_Seguimiento  <int> 400, 4500, 1012, 1925, 1504, 2503, 1832, 2466,
2400, …
$ Estado            <chr> "Fallecido", "Censurado", "Fallecido", "Fallecido",
"…
$ Medicamento       <chr> "D_penicilamina", "D_penicilamina",
"D_penicilamina",…
$ Edad              <int> 21464, 20617, 25594, 19994, 13918, 24201, 20284,
1937…
$ Sexo              <chr> "Mujer", "Mujer", "Hombre", "Mujer", "Mujer",
"Mujer"…
```

Estadísticos descriptivos y otros parámetros para exploración de datos

```{r}
skimr::skim(cirrosis)
```

A tibble: 12 × 11

	skim_variable <chr>	n_missing <int>	complete_rate <dbl>	mean <dbl>
1	ID	0	1.0000000	209.500000
2	Dias_Seguimie…	0	1.0000000	1917.782297
3	Edad	0	1.0000000	18533.351675
4	Bilirrubina	0	1.0000000	3.220813
5	Colesterol	134	0.6794258	369.510563
6	Albumina	0	1.0000000	3.497440
7	Cobre	108	0.7416268	97.648387

Resumen por variable

```r
summary(cirrosis)
```

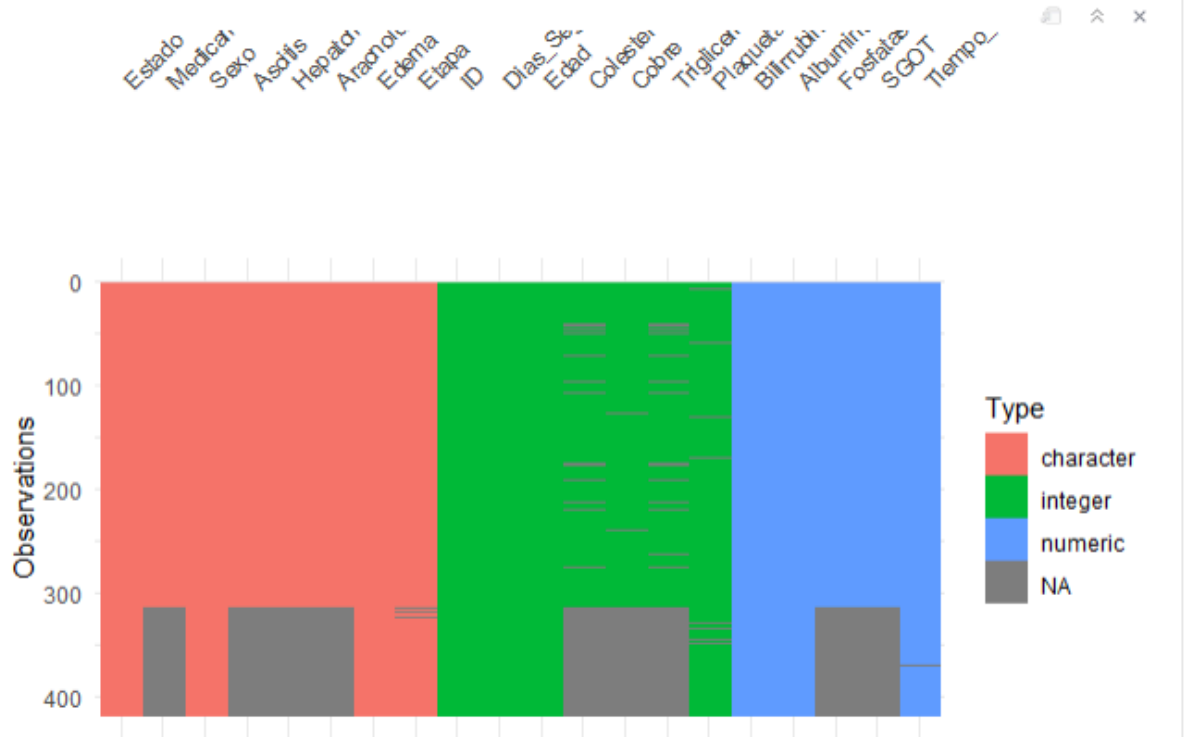
```
       ID         Dias_Seguimiento    Estado
 Min.   :  1.0   Min.   :  41     Length:418
 1st Qu.:105.2   1st Qu.:1093     Class :character
 Median :209.5   Median :1730     Mode  :character
 Mean   :209.5   Mean   :1918
 3rd Qu.:313.8   3rd Qu.:2614
 Max.   :418.0   Max.   :4795

 Medicamento            Edad          Sexo
 Length:418         Min.   : 9598   Length:418
 Class :character   1st Qu.:15644   Class :character
 Mode  :character   Median :18628   Mode  :character
                    Mean   :18533
                    3rd Qu.:21273
                    Max.   :28650
```

Visualmente

```r
View(cirrosis)
```

```r
visdat::vis_dat(cirrosis)
```



DataExplorer::create_report(cirrosis)

Limpieza de datos

Paso uno: corregir los nombres de variables.

Clean names es una función del paquete janitor

```{r}
cirrosis_1 = clean_names(cirrosis)
```

Nota el constraste (la función `names()` imprime los nombres de columnas de un *dataset*)

```{r}
names(cirrosis_1)
```

```
 [1] "id"                 "dias_seguimiento"
 [3] "estado"             "medicamento"
 [5] "edad"               "sexo"
 [7] "ascitis"            "hepatomegalia"
 [9] "aracnoides"         "edema"
[11] "bilirrubina"        "colesterol"
[13] "albumina"           "cobre"
[15] "fosfatasa_alcalina" "sgot"
[17] "trigliceridos"      "plaquetas"
[19] "tiempo_protrombina" "etapa"
```

Paso dos: convertir celdas vacías a NA

```{r}
cirrosis_1 = mutate_if(cirrosis_1, is.character, list(~na_if(.,"")))
```

Paso tres: eliminar columnas o filas vacias.

```{r}
cirrosis_2= remove_empty(cirrosis_1, which = c("rows", "cols"))
```

Optimizando el código

Corregir nombres, celdas vacías a NA y eliminar columnas o filas vacías.

```{r}
cirrosis_1 = cirrosis |>
  clean_names() |>
  mutate_if(is.character, list(~ na_if(., ""))) |>
  remove_empty(which = c("rows", "cols"))
```

Paso 4: corregir errores ortográficos o valores inválidos

Inspección tabular

```{r}
cirrosis |> count(Dias_Seguimiento) # Cambia de variable categórica
```

Description: df [399 × 2]

Dias_Seguimi... <int>	n <int>
41	2
43	1
51	1
71	1
77	1
94	1
110	1
111	1
130	1
131	1

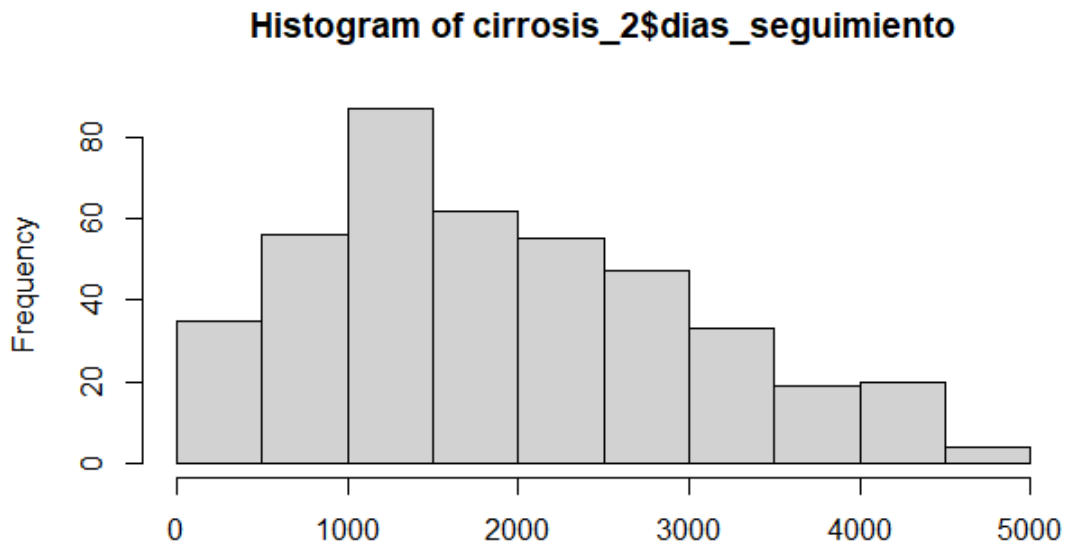
Paso 6: Transformar una variable

Transformación a logaritmo

```{r}
summary(cirrosis_2$dias_seguimiento)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     41    1093    1730    1918    2614    4795 
```

```{r}
hist(cirrosis_2$dias_seguimiento)
```



Transformación a binario

```{r}
data = cirrosis_3 |>
  mutate(
    tiempo_desde_dx_c = case_when(
      edad <  10 ~ "< 10",
      edad >= 10 ~ ">= 10"
    )
  ) |>
  mutate(tiempo_desde_dx_c = factor(tiempo_desde_dx_c, levels = "< 10", ">= 10"
))
```

Transformando valores a valores perdidos usando la función `na_if()`

```{r}
cirrosis_4 = cirrosis_3 |>
  mutate(dias_seguimiento = na_if(dias_seguimiento, -7))
```

Transformando valores a valores perdidos usando la función `case_when()`

```{r}
cirrosis_5 = cirrosis_4 |>
  mutate(ID = case_when(dias_seguimiento %in% c(3, 999) ~ NA,
                        TRUE ~ edad_a))
```

Paso 7: Renombrar una variable

Imprimir los nombres. ¿Cuáles necesitann cambio?

```{r}
names(cirrosis_3)
```

```
 [1] "id"                 "dias_seguimiento"   "estado"
 [4] "medicamento"        "edad"               "sexo"
 [7] "ascitis"            "hepatomegalia"      "aracnoides"
[10] "edema"              "bilirrubina"        "colesterol"
[13] "albumina"           "cobre"              "fosfatasa_alcalina"
[16] "sgot"               "trigliceridos"      "plaquetas"
[19] "tiempo_protrombina" "etapa"
```

Cambiando un nombre de variables

```{r}
cirrosis_3 <- cirrosis_2 |>
  rename( estado = etapa)
```

Comprobando

```{r}
names(cirrosis_3)
```

```{r}
DataExplorer::create_report(cirrosis_4)
```