Unidad 1: Análisis exploratorio de datos

Mg. J. Eduardo Gamboa U.

Table of contents

Carga de paquetes

```
library(readr)
library(dplyr)
library(modeest)
library(sjstats)
library(cleaner)
library(DescTools)
library(moments)
library(reflimR)
library(janitor)
library(summarytools)
library(ggplot2)
library(waffle)
library(treemapify)
library(skimr)
library(explore)
```

Lectura de datos

Se empleará el archivo `Salud.csv`, el cual recopila datos de pacientes en torno a las siguientes cuatro variables:

- Edad (en años)
- Tiempo semanal de ejercicios (en minutos)
- Índice de Masa Corporal
- Presión sistólica (en mmHg)

```
datos <- read_csv('Salud.csv')
```

```
Rows: 100 Columns: 4
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (4): Edad, Minutos_ejercicio, IMC, Presion_sistolica
```

```
i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
datos |> head(5)
```

```
# A tibble: 5 x 4
   Edad Minutos_ejercicio   IMC Presion_sistolica
  <dbl>             <dbl> <dbl>             <dbl>
1    31               267  15.9               111
2    42               142  20.1               142
3    61                58  20.2               139
4    41                25  17.8               120
5    40                46  17.4               133
```

```
datos |> tail(3)
```

```
# A tibble: 3 x 4
   Edad Minutos_ejercicio   IMC Presion_sistolica
  <dbl>             <dbl> <dbl>             <dbl>
1    44               176  20.4               111
2    37                 8  20.1               131
3    44                98  15.7               128
```

Medidas de tendencia central

Media

Ejemplo 1

Interpretar la media aritmética de la edad

```
datos |> summarize(Media = mean(Edad))
```

```
# A tibble: 1 x 1
  Media
  <dbl>
1  46.0
```

La edad promedio de los pacientes es de 46 años.

Ejemplo 2

Interpretar la presión sistólica media de los pacientes mayores de 50 años.

```
datos |> filter(Edad > 50) |> summarize(Media = mean(Presion_sistolica))
```

```
# A tibble: 1 x 1
  Media
  <dbl>
1  130.
```

La presión sistólica promedio de los pacientes mayores de 50 años es de 130 mmHg.

Mediana

Ejemplo 3

Interpretar la mediana del IMC

```
datos |> summarize(Mediana = median(IMC))
```

```
# A tibble: 1 x 1
  Mediana
    <dbl>
1    19.2
```

Al menos la mitad de las personas tiene un IMC menor o igual a 19.2.

Ejemplo 4

Interpretar la mediana de la presión sistólica para las personas que son sedentarias (menos de 30 minutos de ejercicios a la semana) y las que no lo son.

```
datos |>
  mutate(Sedentario = ifelse(Minutos_ejercicio<30,"Sí","No")) -> datos

datos |>
  group_by(Sedentario) |>
  summarize(Medianas = median(Presion_sistolica))
```

```
# A tibble: 2 x 2
  Sedentario Medianas
  <chr>         <dbl>
1 No              121
2 Sí              138.
```

Al menos la mitad de las personas sedentarias presenta una presión sistólica de como máximo 138 mmHg (¡elevada!). Por otro lado, al menos el 50% de las personas que no son sedentarias tiene una presión sistólica menor o igual a 121 mmHg (casi en el rango normal).

Moda

Ejemplo 5

Interpretar la moda de la presión sistólica

```
datos |>
  summarize(Moda = mfv(Presion_sistolica))
```

```
# A tibble: 1 x 1
   Moda
  <dbl>
1   121
```

La presión sistólica más frecuente es de 121 mHg.

Ejemplo 6

Interpretar la moda de la edad

```
datos |>
  reframe(Moda = mfv(Edad))
```

```
# A tibble: 2 x 1
   Moda
  <dbl>
1    36
2    59
```

Las edades más frecuentes de los pacientes son 36 y 59 años.

Ejemplo 7

Intepretar la moda del tiempo semanal de ejercicio de los pacientes sedentarios

```
datos |>
  filter(Sedentario == "Sí") |>
  reframe(Moda = mfv(Minutos_ejercicio))
```

```
# A tibble: 1 x 1
   Moda
  <dbl>
1    15
```

El tiempo de ejercicios más frecuente entre los pacientes sedentarios es de 15 minutos.

Medidas de posición

Ejemplo 8

Interpretar el percentil 41 de la edad

```
datos |>
  summarize(P41 = quantile(Edad, 0.41))
```

```
# A tibble: 1 x 1
    P41
  <dbl>
1    42
```

Al menos el 41% de los pacientes tiene 42 años de edad o menos.

Ejemplo 9

Interpretar los percentiles 12 y 74 de los tiempos semanales de ejercicio de las personas no sedentarias

```
datos |>
  filter(Sedentario == "No") |>
  reframe(Percentiles = quantile(Minutos_ejercicio, c(0.12,0.74)))
```

```
# A tibble: 2 x 1
  Percentiles
        <dbl>
1        61.4
2       241.
```

Al menos el 12% de los pacientes no sedentarios realiza como máximo 61.4 minutos de ejercicio a la semana, mientras que al menos el 74% realiza hasta 241 minutos semanales de actividad física.

Ejemplo 10

Interpretar los cuartiles del IMC de las personas adultas mayores (60 años a más)

```
datos |>
  filter(Edad >= 60) |>
  reframe(Cuartiles = quantile(IMC, c(0.25,0.50,0.75)))
```

```
# A tibble: 3 x 1
  Cuartiles
      <dbl>
1      19.4
2      20
3      20.9
```

Al menos el 25% de los pacientes tiene un IMC igual o inferior a 19.4, mientras que como máximo el 50% tiene un IMC igual o inferior a 20. Además, hasta el 75% de los pacientes presenta un IMC igual o inferior a 20.9.

Ejemplo 11

¿Cuál es el tiempo máximo de ejercicio semanal que realiza un paciente joven (menor de 30 años) para estar dentro del 20% que menos ejercicio realiza?

```
datos |>
  filter(Edad < 30) |>
  summarize(P20 = quantile(Minutos_ejercicio, 0.20))
```

```
# A tibble: 1 x 1
    P20
  <dbl>
1    58
```

58 minutos semanales es el tiempo máximo de ejercicio que realiza un paciente joven (menor de 30 años) para estar dentro del 20% que menos ejercicio realiza.

Medidas de dispersión

Rango

Ejemplo 12

Interpretar el rango de la edad

```
datos |> summarize(r = Range(Edad))
```

```
# A tibble: 1 x 1
      r
  <dbl>
1    49
```

La amplitud de la edad es de 49 años.

Ejemplo 13

Interpretar el rango del IMC para cada grupo de personas según su nivel de actividad física (sedentario / no sedentario).

```
datos |> group_by(Sedentario) |> summarize(r = Range(IMC))
```

```
# A tibble: 2 x 2
  Sedentario     r
  <chr>      <dbl>
1 No           9.5
2 Sí           4.7
```

La amplitud del IMC de las personas sedentarias es de 4.7 puntos, mientras que para las no sedentarias es de 9.5 puntos.

Rango intercuartil

Ejemplo 14

Interpretar el rango intercuartil de la edad

```
datos |> reframe(Q = quantile(Edad, c(0.25,0.75)))
```

```
# A tibble: 2 x 1
      Q
  <dbl>
1  36
2  58.2
```

```
datos |> summarize(ric = IQR(Edad))
```

```
# A tibble: 1 x 1
    ric
  <dbl>
1  22.2
```

La amplitud del 50% central de las edades es de 22.2 años.

Ejemplo 15

Interpretar el rango intercuartil del IMC para cada grupo de personas según su nivel de actividad física (sedentario / no sedentario).

```
datos |> group_by(Sedentario) |> summarize(ric = IQR(IMC))
```

```
# A tibble: 2 x 2
  Sedentario   ric
  <chr>      <dbl>
1 No           2.6
2 Sí           1.7
```

La amplitud del 50% central de los datos de IMC de las personas sedentarias es de 1.7 puntos, mientras que para las no sedentarias es de 2.6 puntos.

Varianza

Ejemplo 16

Calcular la varianza del tiempo semanal de ejercicios.

```
datos |> summarize(s2 = var(Minutos_ejercicio))
```

```
# A tibble: 1 x 1
     s2
  <dbl>
1 8183.
```

La varianza del tiempo semanal de ejercicios es de 8183 $minutos^2$.

Desviación estándar

Ejemplo 17

Interpretar la desviación estándar del tiempo semanal de ejercicios.

```
datos |> summarize(s = sd(Minutos_ejercicio))
```

```
# A tibble: 1 x 1
      s
  <dbl>
1  90.5
```

```
datos |> summarize(s = sqrt(var(Minutos_ejercicio)))
```

```
# A tibble: 1 x 1
      s
  <dbl>
1  90.5
```

En promedio, el tiempo semanal de ejercicios se desvía 90.5 minutos respecto su media.

Coeficiente de variabilidad

Ejemplo 18

¿Qué variable presenta mayor variabilidad: el IMC o la presión sistólica?

```
datos |> summarize(s_imc   = sd(IMC),
                   s_pres  = sd(Presion_sistolica),
                   m_imc   = mean(IMC),
                   m_pres  = mean(Presion_sistolica),
                   cv_imc  = cv(IMC)*100,
                   cv_pres = cv(Presion_sistolica)*100)
```

```
# A tibble: 1 x 6
  s_imc s_pres m_imc m_pres cv_imc cv_pres
  <dbl>  <dbl> <dbl>  <dbl>  <dbl>   <dbl>
1  2.03   12.7  19.1   124.   10.7    10.3
```

El IMC presenta mayor variabilidad (cv = 10.7%) que la presión sistólica (cv = 10.3%).

Ejemplo 19

Los pacientes se dividen en 3 grupos: joven (menor de 30 años),adulto (de 30 a 59 años) y adulto mayor (de 60 a más años). ¿En qué grupo se observa mayor variabilidad en el tiempo semanal de ejercicios?

```
datos |>  mutate(Grupo_Edad = case_when(Edad < 30 ~ "Joven",
                                        Edad >= 30 & Edad < 60 ~ "Adulto",
                                        Edad >= 60 ~ "Adulto mayor")) |>
  group_by(Grupo_Edad) |>
  summarise(s  = sd(Minutos_ejercicio),
            m  = mean(Minutos_ejercicio),
            cv = cv(Minutos_ejercicio))
```

```
# A tibble: 3 x 4
  Grupo_Edad       s     m    cv
  <chr>        <dbl> <dbl> <dbl>
1 Adulto        89.6  167. 0.536
2 Adulto mayor  84.4  114. 0.740
3 Joven         91.8  156. 0.588
```

El grupo con mayor variabilidad en el tiempo semanal de ejercicios es el de adultos mayores.

Medidas de asimetría

Coeficiente de asimetría de Fisher Pearson

Ejemplo 20

Interpretar el coeficiente de asimetría de Fisher Pearson para cada variable.

```
datos |>
   summarise(across(where(is.numeric), ~ skewness(.x)))
```

```
# A tibble: 1 x 4
    Edad Minutos_ejercicio    IMC Presion_sistolica
   <dbl>             <dbl>  <dbl>             <dbl>
1 -0.134            -0.174 -0.318             0.165
```

Las cuatro variables presentan distribuciones que tienden a la simetría, siendo la edad la que más se acerca a la simetría y el IMC la que más se aleja.

Coeficiente de asimetría de Bowley

Ejemplo 21

Interpretar el coeficiente de asimetría de Bowley para cada variable.

```
datos |>
   summarise(across(where(is.numeric), ~ bowley(.x)))
```

```
# A tibble: 1 x 4
   Edad Minutos_ejercicio    IMC Presion_sistolica
  <dbl>             <dbl>  <dbl>             <dbl>
1 0.191           -0.0781 0.0196             0.160
```

Las cuatro variables presentan distribuciones que tienden a la simetría en el 50% central de los datos, siendo el IMC la que más se acerca a la simetría y la edad la que más se aleja. ¿Es una contradicción respecto a la interpretación con el coeficiente de asimetría de Fisher Pearson?

Tablas de frecuencia

Tablas de frecuencia para variables cualitativas

```
datos |>
  pull(Sedentario) |>
  tabyl() |>
  adorn_totals("row")
```

```
 pull(datos, Sedentario)   n percent
                      No  88    0.88
                      Sí  12    0.12
                   Total 100    1.00
```

Tablas de frecuencia para variables cuantitativas discretas

Siempre que su rango de valores no sea extenso, de lo contrario tratar como variable cuantitativa continua:

```
datos |>
  filter(Edad>=60) |>
  summarytools::freq(Edad)
```

```
Frequencies
filter(.data = datos, Edad >= 60)$Edad
Type: Numeric

              Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
---------- ------ --------- -------------- --------- --------------
        61      2     10.00          10.00     10.00          10.00
        62      4     20.00          30.00     20.00          30.00
        63      3     15.00          45.00     15.00          45.00
        64      3     15.00          60.00     15.00          60.00
        65      1      5.00          65.00      5.00          65.00
        66      3     15.00          80.00     15.00          80.00
        68      3     15.00          95.00     15.00          95.00
        69      1      5.00         100.00      5.00         100.00
      <NA>      0                                0.00         100.00
     Total     20    100.00         100.00    100.00         100.00
```

Tablas de frecuencia para variables cuantitativas continuas

```
datos |>
  pull(Edad) |>
  DescTools::Freq()
```

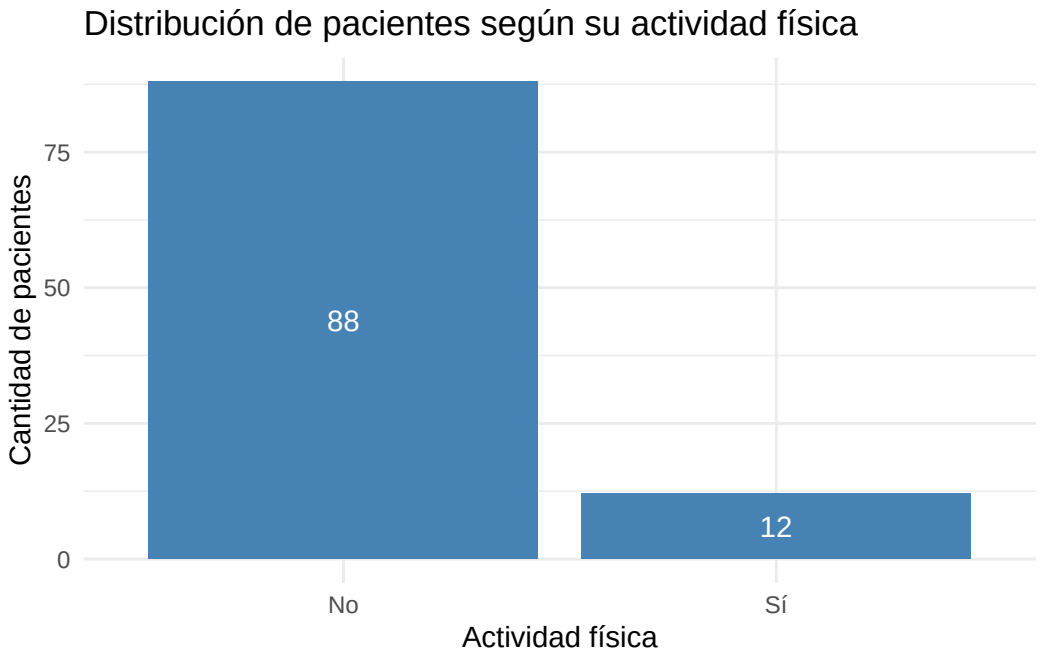
```
      level  freq   perc  cumfreq  cumperc
1   [20,25]     8   8.0%        8     8.0%
2   (25,30]    10  10.0%       18    18.0%
3   (30,35]     6   6.0%       24    24.0%
4   (35,40]    14  14.0%       38    38.0%
5   (40,45]    13  13.0%       51    51.0%
6   (45,50]     7   7.0%       58    58.0%
7   (50,55]    11  11.0%       69    69.0%
8   (55,60]    11  11.0%       80    80.0%
9   (60,65]    13  13.0%       93    93.0%
10  (65,70]     7   7.0%      100   100.0%
```

Gráficas

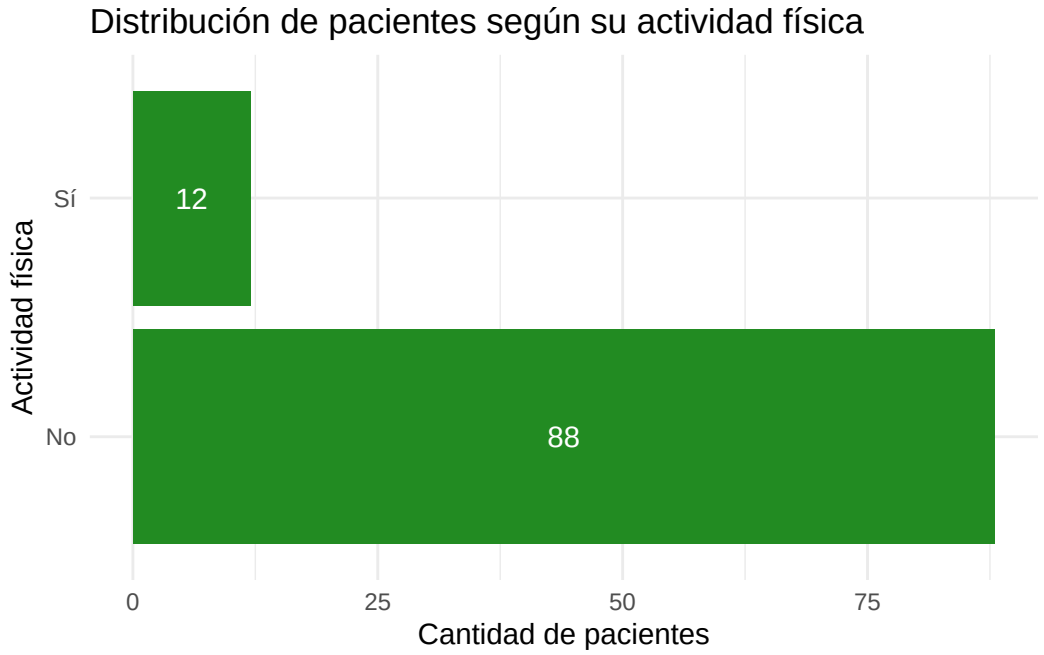
Gráficas para variables cualitativas

Gráfico de barras

```
datos |>
  count(Sedentario) |>
  ggplot(aes(x = Sedentario, y = n)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = n, y = n / 2), color = "white") +
  labs(title = "Distribución de pacientes según su actividad física",
       x = "Actividad física",
       y = "Cantidad de pacientes") +
  theme_minimal()
```



```r
datos |>
  count(Sedentario) |>
  ggplot(aes(x = Sedentario, y = n)) +
  geom_bar(stat = "identity", fill = "forestgreen") +
  geom_text(aes(label = n, y = n / 2), color = "white") +
  labs(title = "Distribución de pacientes según su actividad física",
       x = "Actividad física",
       y = "Cantidad de pacientes") +
  coord_flip()+
  theme_minimal()
```



```r
datos |>
  ggplot(aes(x = Sedentario)) +
  geom_bar(fill = "darkblue") +
  labs(title = "Distribución de pacientes según su actividad física",
       x = "Actividad física",
       y = "Cantidad de pacientes") +
  theme_minimal()
```

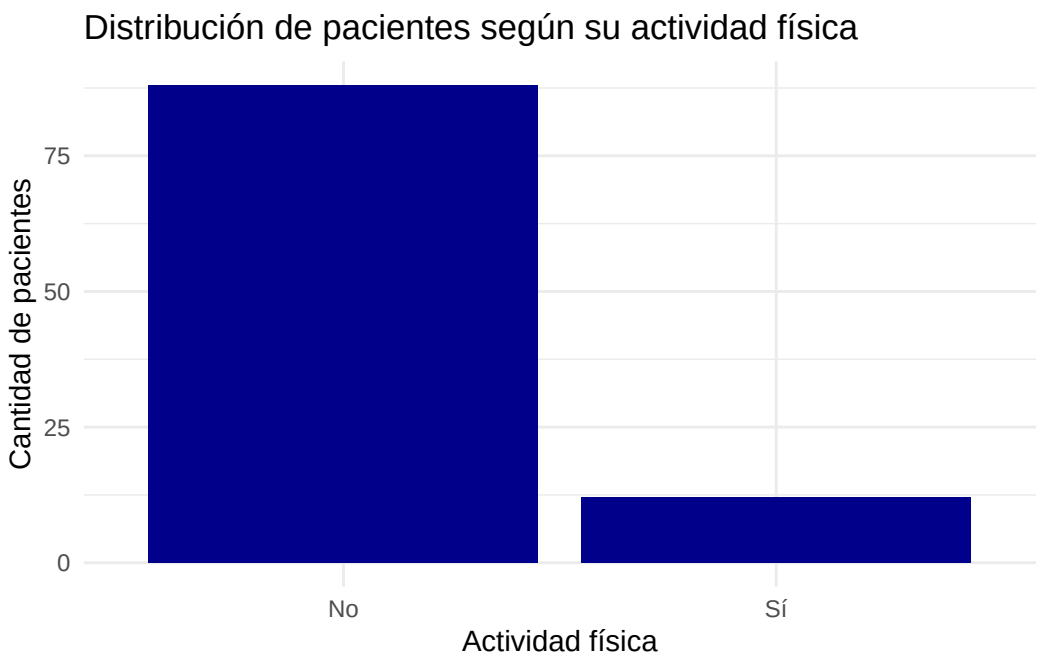


Gráfico circular

```
datos |>
  count(Sedentario) |>
  mutate(porcentaje = round(n / sum(n) * 100, 1),
         etiqueta = paste0(Sedentario, ": ", porcentaje, "%")) |>
  ggplot(aes(x = "", y = n, fill = Sedentario)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start = 0) +
  geom_text(aes(label = etiqueta), position = position_stack(vjust = 0.5)) +
  labs(title = "Distribución de pacientes según su actividad física") +
  theme_void()  # elimina ejes
```

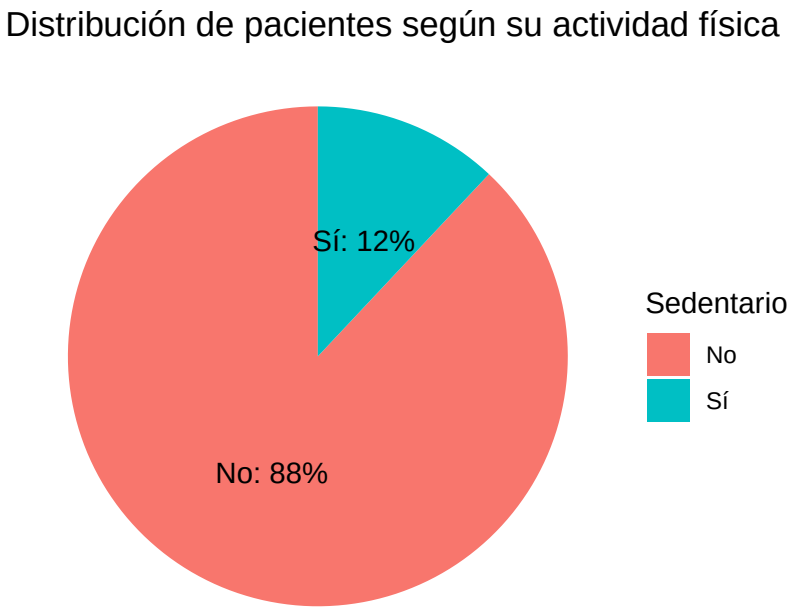


Gráfico waffle

```
datos |>
  count(Sedentario) |>
  mutate(n = round(n / sum(n) * 100)) |>
  pull(n) |>
  waffle(rows = 5, title = "Distribución de pacientes")
```

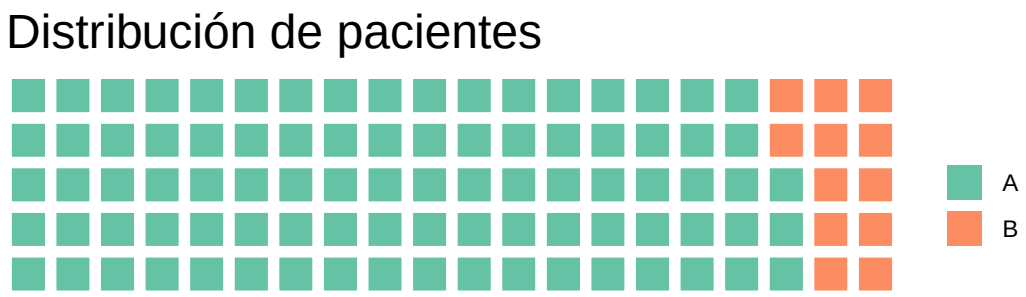


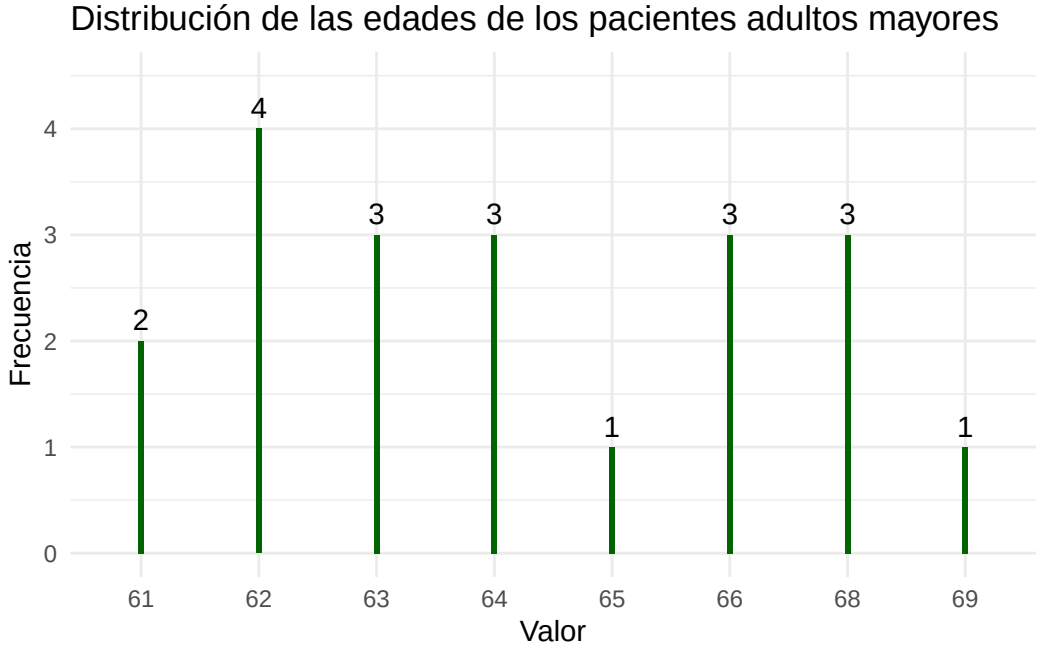
Gráfico treemapify

```
datos |>
  count(Sedentario) |>
  ggplot(aes(area = n, fill = Sedentario, label = Sedentario)) +
  geom_treemap() +
  geom_treemap_text(colour = "white", place = "centre", grow = TRUE) +
  labs(title = "Distribución de pacientes") +
  theme_minimal()
```

Gráficas para variables cuantitativas

Gráfica de varas (barras)

```
datos |>
  filter(Edad > 60) |>
  count(Edad) |>
  ggplot(aes(x = factor(Edad), y = n)) +
  geom_bar(stat = "identity", fill = "darkgreen", width = 0.05) +
  geom_text(aes(label = n), vjust = -0.5) +
  scale_y_continuous(limits = c(0,4.5)) +
  labs(title = "Distribución de las edades de los pacientes adultos mayores",
       x = "Valor",
       y = "Frecuencia") +
  theme_minimal()
```



Histograma

```
datos |>
  ggplot(aes(x = Edad)) +
  geom_histogram(binwidth = 5, fill = "mediumorchid", color = "white") +
  labs(title = "Distribución de la edad", x = "Edad", y = "Frecuencia") +
  theme_minimal()
```

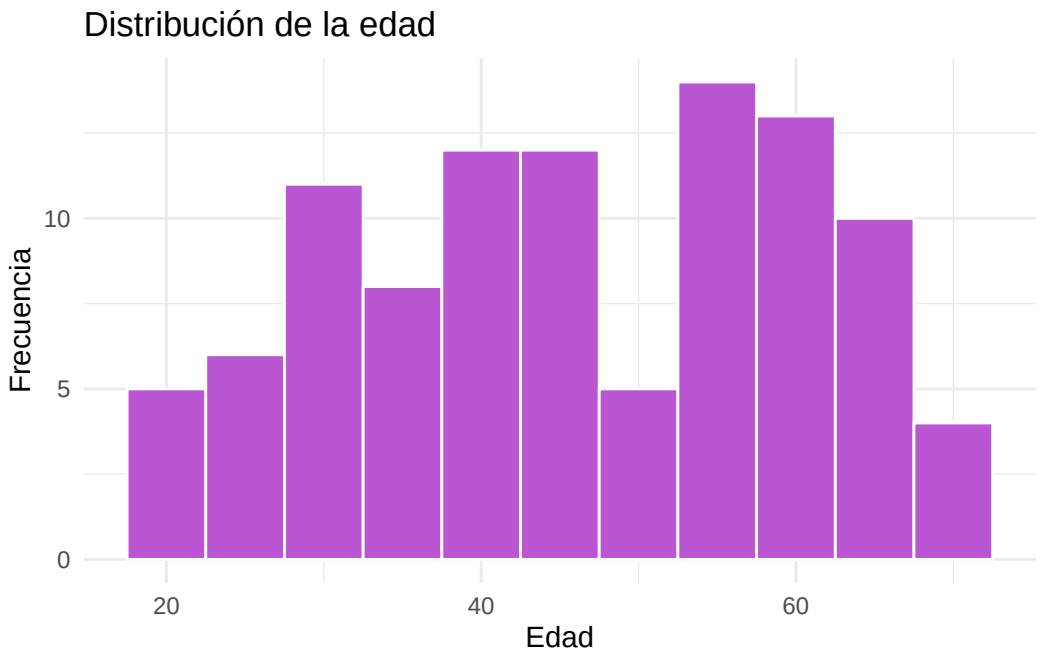
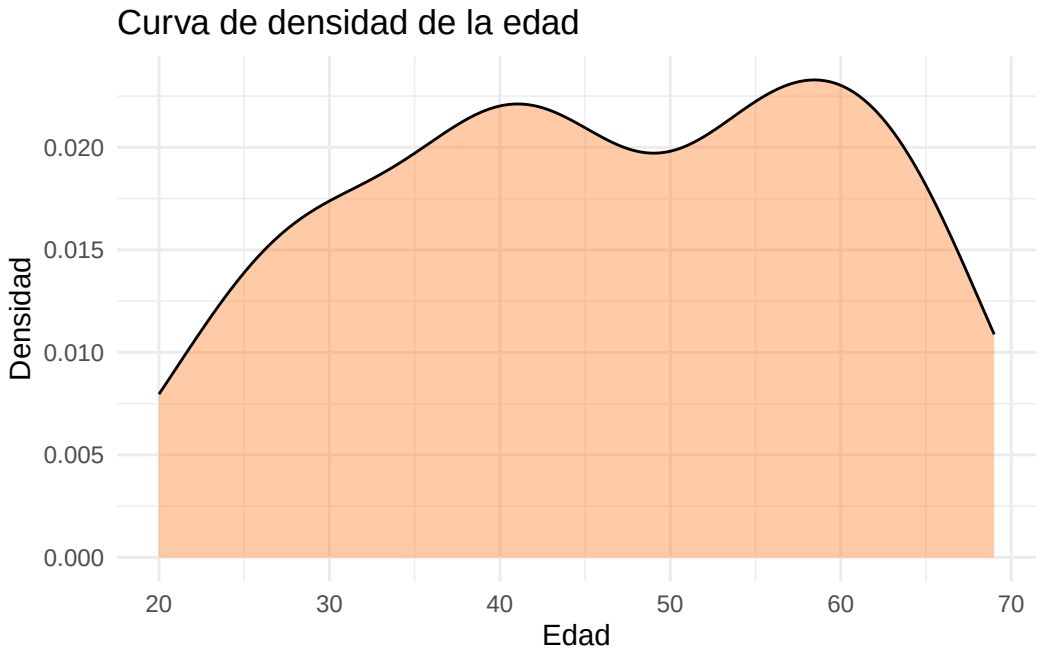


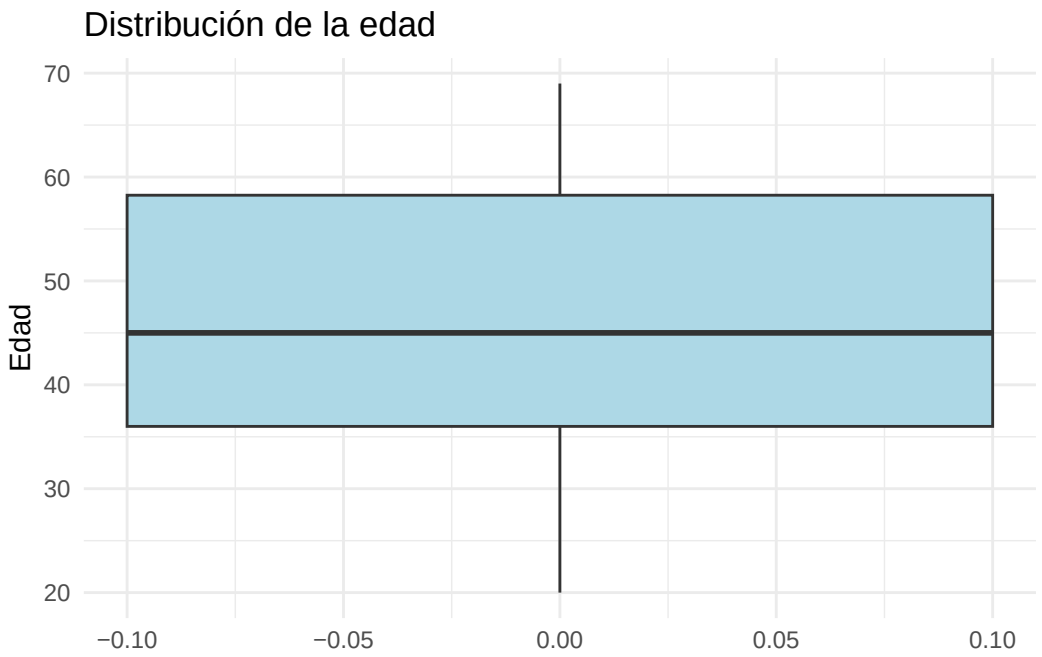
Gráfico de densidad

```
datos |>
  ggplot(aes(x = Edad)) +
  geom_density(fill = "chocolate1", alpha = 0.4) +
  labs(title = "Curva de densidad de la edad", x = "Edad", y = "Densidad") +
  theme_minimal()
```

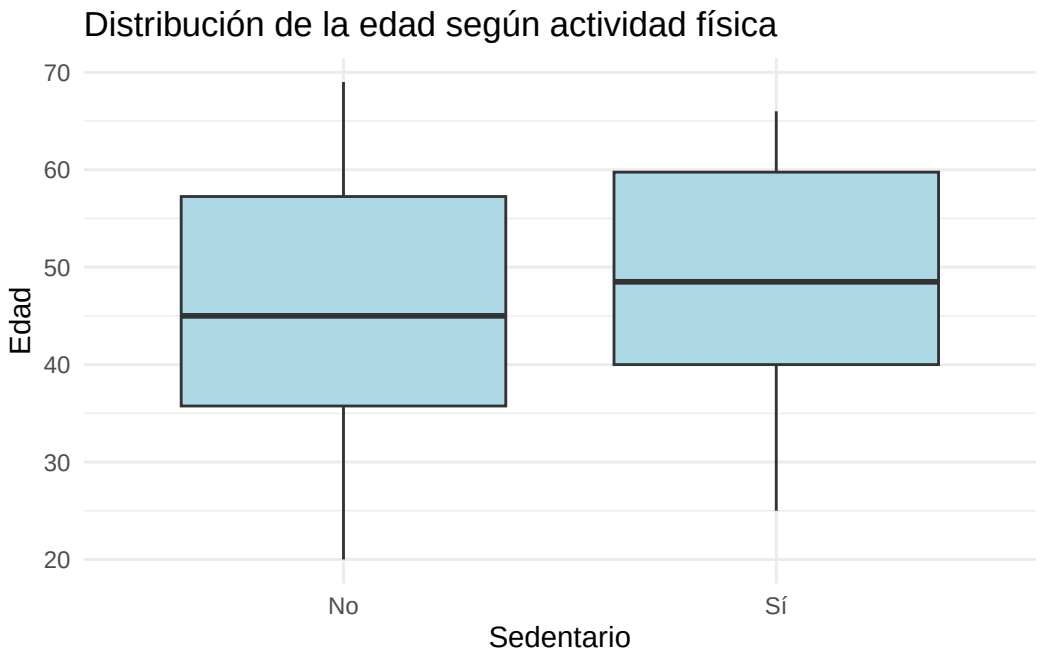


Boxplot

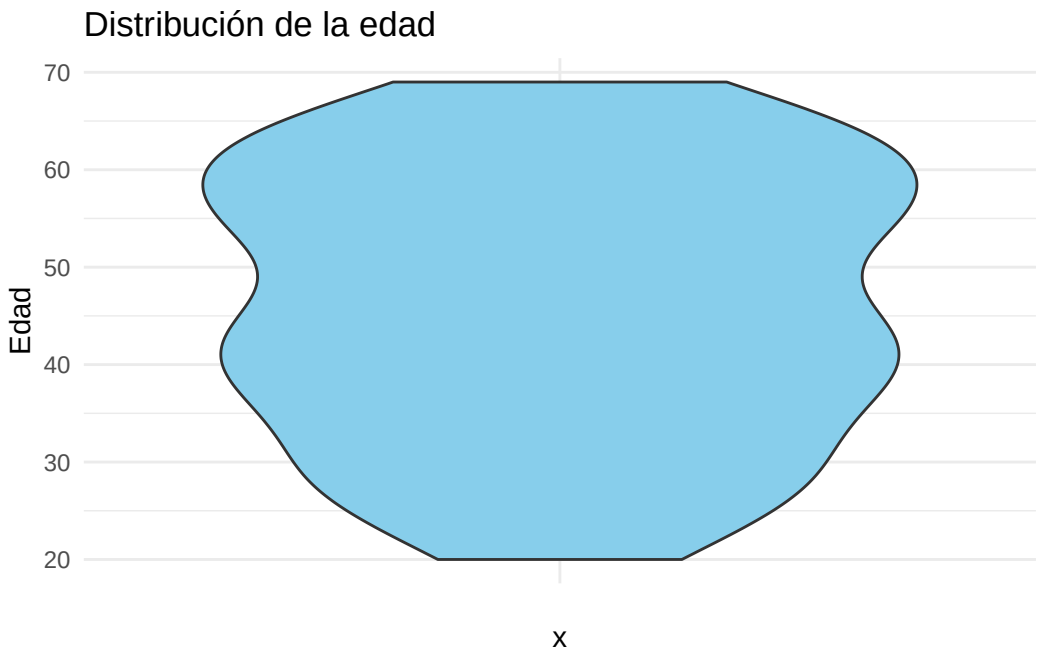
```
datos |>
  ggplot(aes(y = Edad)) +
  geom_boxplot(fill = "lightblue", width = 0.2) +
  labs(title = "Distribución de la edad", y = "Edad") +
  theme_minimal()
```



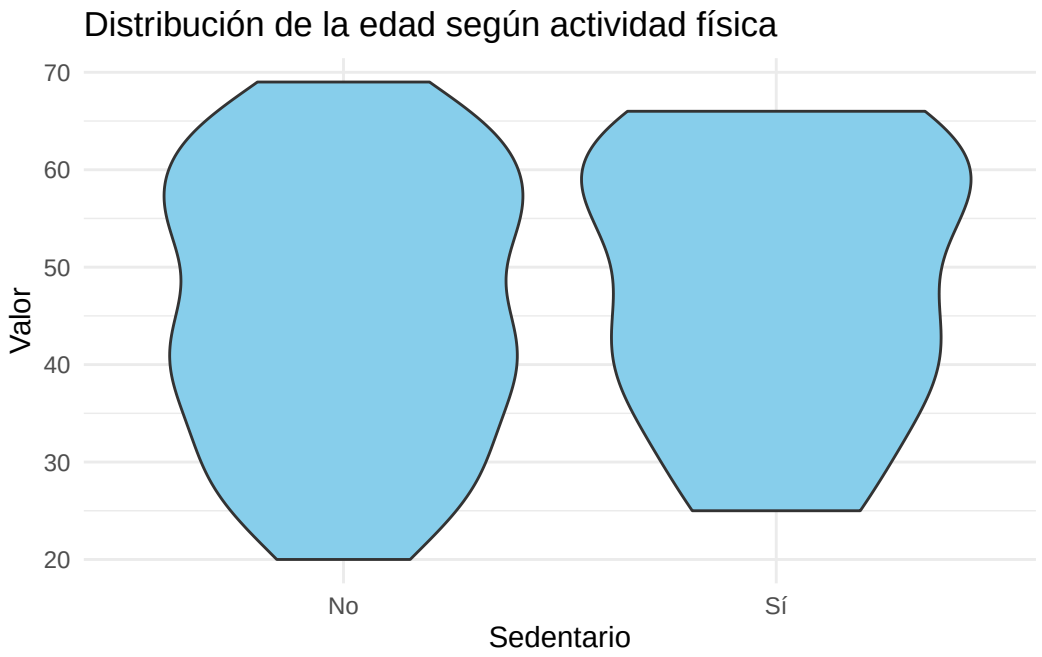
```r
datos |>
  ggplot(aes(y = Edad, x = Sedentario)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Distribución de la edad según actividad física", y = "Edad") +
  theme_minimal()
```



```r
datos |>
  ggplot(aes(x = "", y = Edad)) +
  geom_violin(fill = "skyblue") +
  labs(title = "Distribución de la edad", y = "Edad") +
  theme_minimal()
```



```r
datos |>
  ggplot(aes(x = Sedentario, y = Edad)) +
  geom_violin(fill = "skyblue") +
  labs(title = "Distribución de la edad según actividad física", y = "Valor") +
  theme_minimal()
```



Resúmenes

```
datos |> skim()
```

Table 1: Data summary

Name	datos
Number of rows	100
Number of columns	5
_____	
Column type frequency:	
character	1
numeric	4
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Sedentario	0	1	2	2	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Edad	0	1	45.97	13.88	20.0	36.00	45.0	58.25	69.0	
Minutos_ejercicio	0	1	154.71	90.46	2.0	69.25	159.0	235.75	294.0	
IMC	0	1	19.07	2.03	13.8	17.95	19.2	20.50	23.3	
Presion_sistolica	0	1	123.58	12.71	100.0	113.00	121.5	133.25	152.0	

```
datos |> group_by(Sedentario) |> skim()
```

Table 4: Data summary

Name	group_by(datos, Sedentari…
Number of rows	100
Number of columns	5
_____	
Column type frequency:	
numeric	4
_____	
Group variables	Sedentario

Variable type: numeric

skim_variable	Sedentario	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Edad	No	0	1	45.67	13.95	20.0	35.75	45.0	57.25	69.0	
Edad	Sí	0	1	48.17	13.73	25.0	40.00	48.5	59.75	66.0	
Minutos_ejercicio	No	0	1	174.18	78.15	33.0	106.75	179.0	242.50	294.0	
Minutos_ejercicio	Sí	0	1	11.92	8.25	2.0	5.50	11.5	15.00	28.0	
IMC	No	0	1	18.88	2.04	13.8	17.70	19.0	20.30	23.3	
IMC	Sí	0	1	20.47	1.35	17.8	19.53	20.6	21.23	22.5	
Presion_sistolica	No	0	1	121.93	12.35	100.0	113.00	121.0	131.25	152.0	
Presion_sistolica	Sí	0	1	135.67	8.17	120.0	130.00	137.5	140.50	148.0	

```
# datos |> explore::explore()
```