Entrega 1 - Notebook Reproducible

[Integrantes del Grupo]

19 Sep 2025

Contents

0.1 1. Carga y diagnóstico inicial

```
ruta_datos <- if (file.exists("data/health_lifestyle_classification.csv")) {
  "data/health_lifestyle_classification.csv"
} else {
  file.path("..", "data", "health_lifestyle_classification.csv")
}

datos <- readr::read_csv(ruta_datos, show_col_types = FALSE) |>
  janitor::clean_names()

datos_resumen <- tibble(
  registros = nrow(datos),
  variables = ncol(datos),
  categ = sum(map_chr(datos, ~ class(.x)[1]) %in% c("character", "factor", "ordered")),
  cuant = sum(map_chr(datos, ~ class(.x)[1]) %in% c("numeric", "integer", "double"))
)

render_table(datos_resumen, caption = "Dimensión y tipología de la base")
```

Table 1: Dimensión y tipología de la base

registros	variables	categ	cuant
1e+05	48	18	30

```
missing_tbl <- datos |>
  summarise(across(everything(), ~ sum(is.na(.)))) |>
  pivot_longer(everything(), names_to = "variable", values_to = "faltantes") |>
  mutate(pct = round(faltantes / nrow(datos), 4)) |>
  arrange(desc(faltantes))

render_table(head(missing_tbl, 15), caption = "Variables con mayor cantidad de NA")
```

Table 2: Variables con mayor cantidad de NA

variable	faltantes	pct
insulin	15836	0.1584

variable	faltantes	pct
heart_rate	14003	0.1400
alcohol_consumption	13910	0.1391
gene_marker_flag	10474	0.1047
income	8470	0.0847
daily_steps	8329	0.0833
blood_pressure	7669	0.0767
survey_code	0	0.0000
age	0	0.0000
gender	0	0.0000
height	0	0.0000
weight	0	0.0000
bmi	0	0.0000
bmi_estimated	0	0.0000
bmi_scaled	0	0.0000

Insight: la mayoría de las variables no tiene valores faltantes; los mayores porcentajes se concentran en `insulin`, `daily_steps`, `exercise_type` y `alcohol_consumption`, por lo que cualquier modelado futuro debe considerar imputación o análisis segmentado.

0.2 2. Variables categóricas

```
cat_vars <- datos |> select(where(~ is.character(.x) || inherits(.x, "factor") || inherits(.x, "ordered
cat_vars_principales <- intersect(c("gender", "target", "education_level", "diet_type"), cat_vars)
```

0.2.1 2.1 Tablas de frecuencia

```
frecuencias_cat <- datos |>
  select(all_of(cat_vars_principales)) |>
  pivot_longer(everything(), names_to = "variable", values_to = "categoria") |>
  filter(!is.na(categoria)) |>
  count(variable, categoria) |>
  group_by(variable) |>
  mutate(pct = round(n / sum(n), 3))

render_table(frecuencias_cat, caption = "Frecuencias y proporciones por categoría")
```

Table 3: Frecuencias y proporciones por categoría

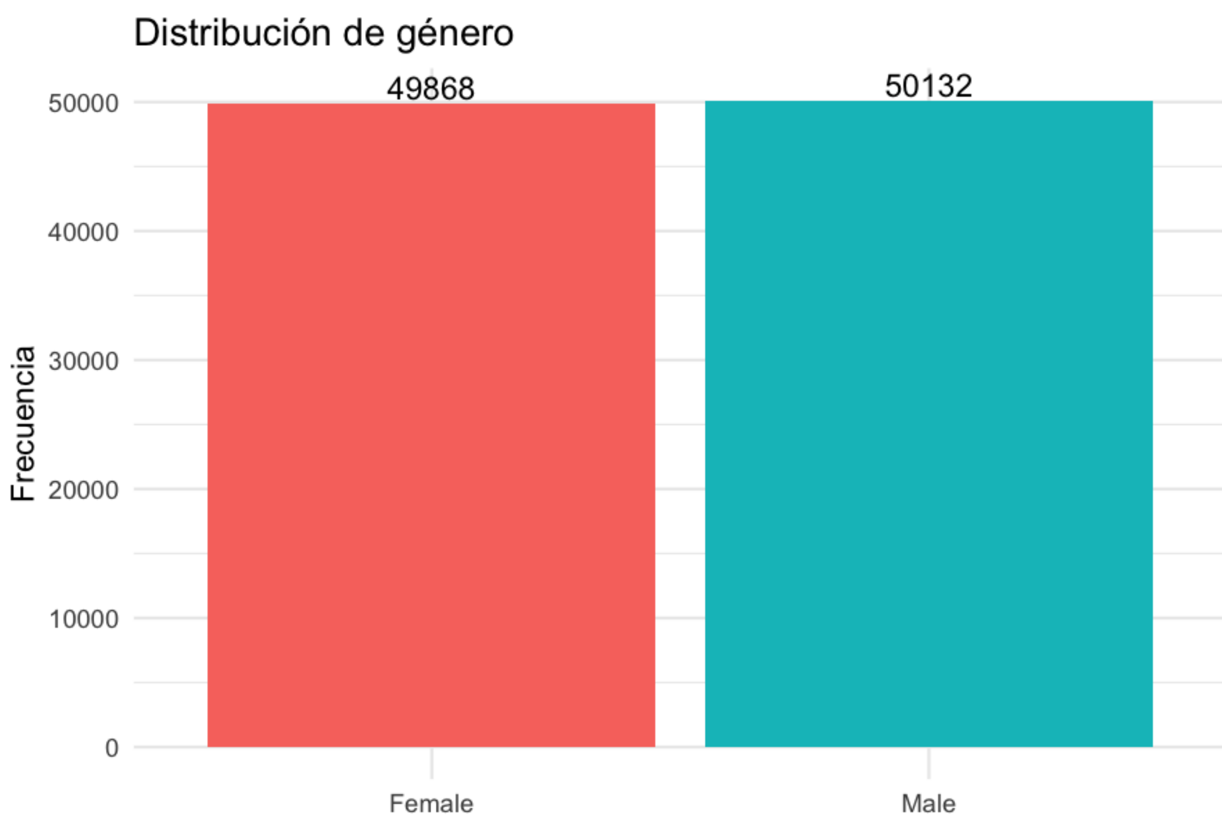
variable	categoria	n	pct
diet_type	Keto	24764	0.248
diet_type	Omnivore	25089	0.251
diet_type	Vegan	25122	0.251
diet_type	Vegetarian	25025	0.250
education_level	Bachelor	25363	0.254
education_level	High School	25028	0.250
education_level	Master	24992	0.250
education_level	PhD	24617	0.246
gender	Female	49868	0.499
gender	Male	50132	0.501
target	diseased	29903	0.299

variable	categoria	n	pct
target	healthy	70097	0.701

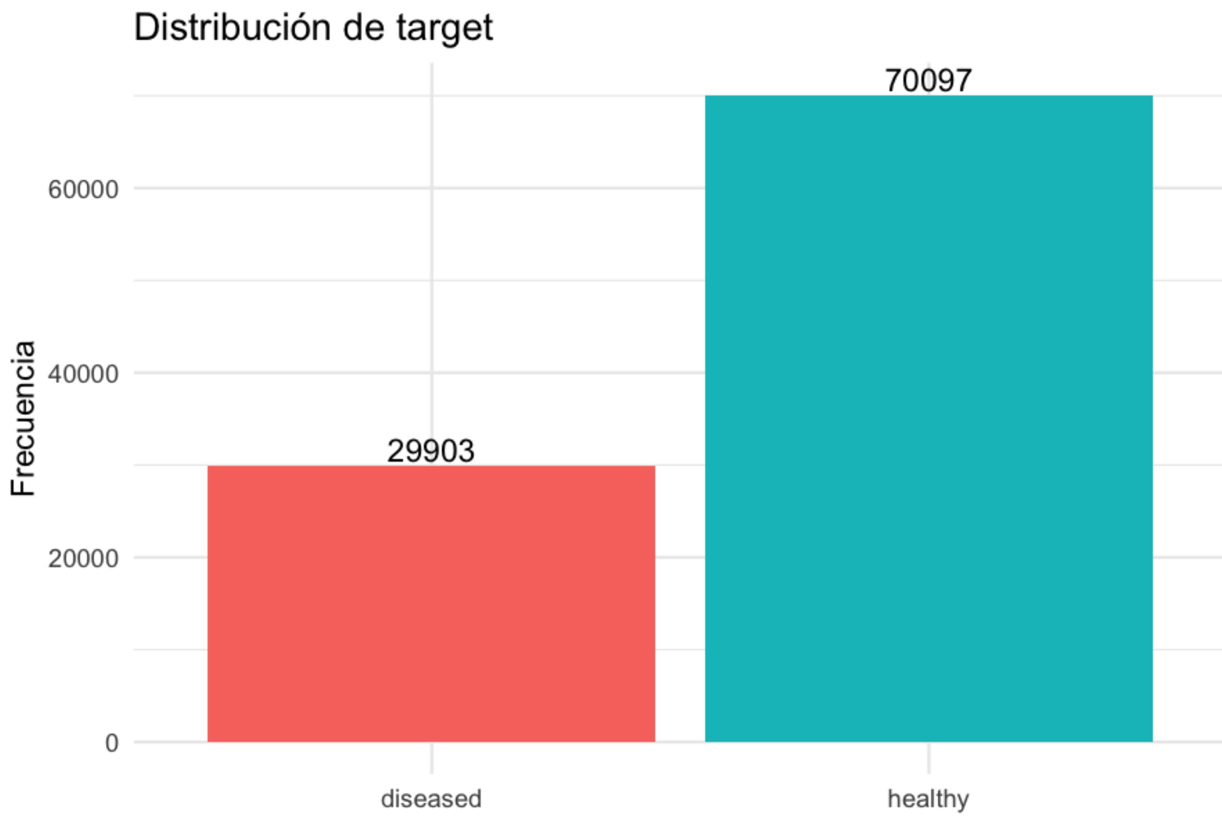
Insight: `target` muestra distribución 70% healthy vs. 30% diseased; `gender` y `education_level` están relativamente balanceadas con ligera concentración en niveles altos de educación.

0.2.2 2.2 Gráficos univariados

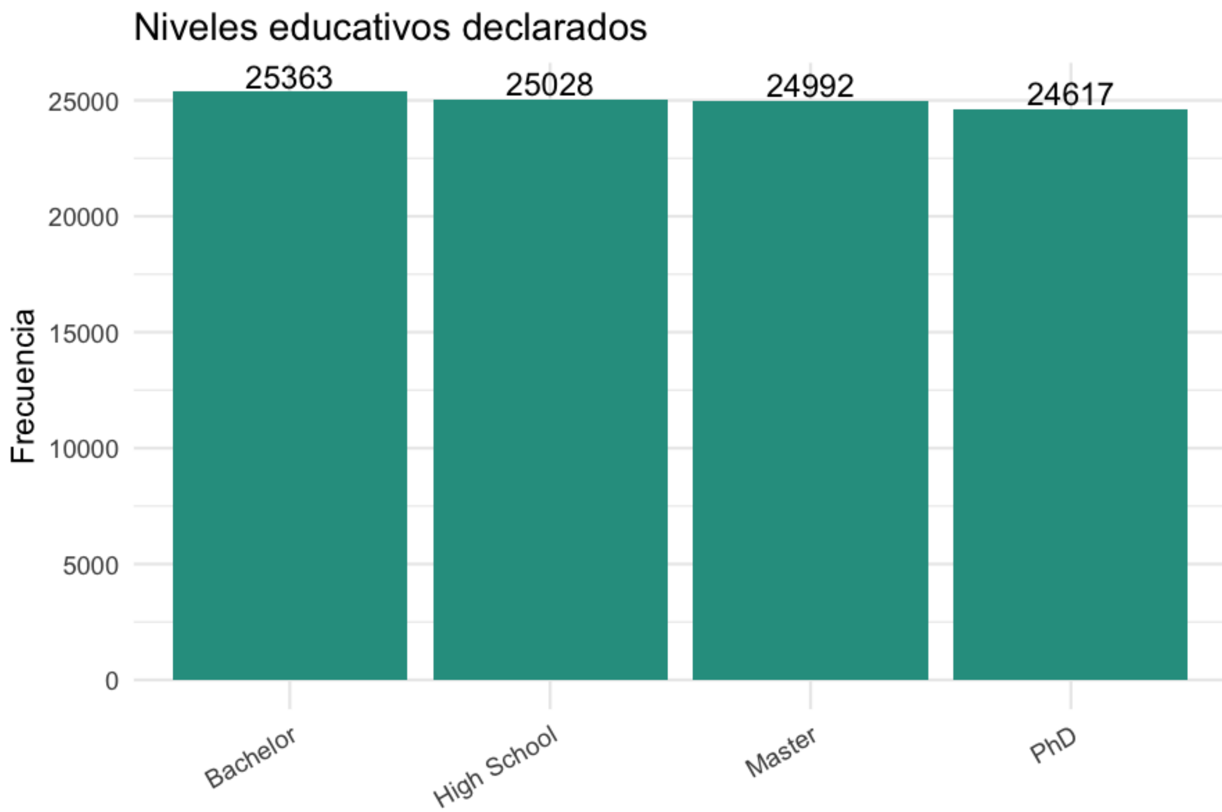
```
plot_gender <- datos |>
  filter(!is.na(gender)) |>
  count(gender) |>
  ggplot(aes(x = gender, y = n, fill = gender)) +
  geom_col(show.legend = FALSE) +
  geom_text(aes(label = n), vjust = -0.2) +
  labs(title = "Distribución de género", x = NULL, y = "Frecuencia") +
  theme_minimal()
plot_gender
```



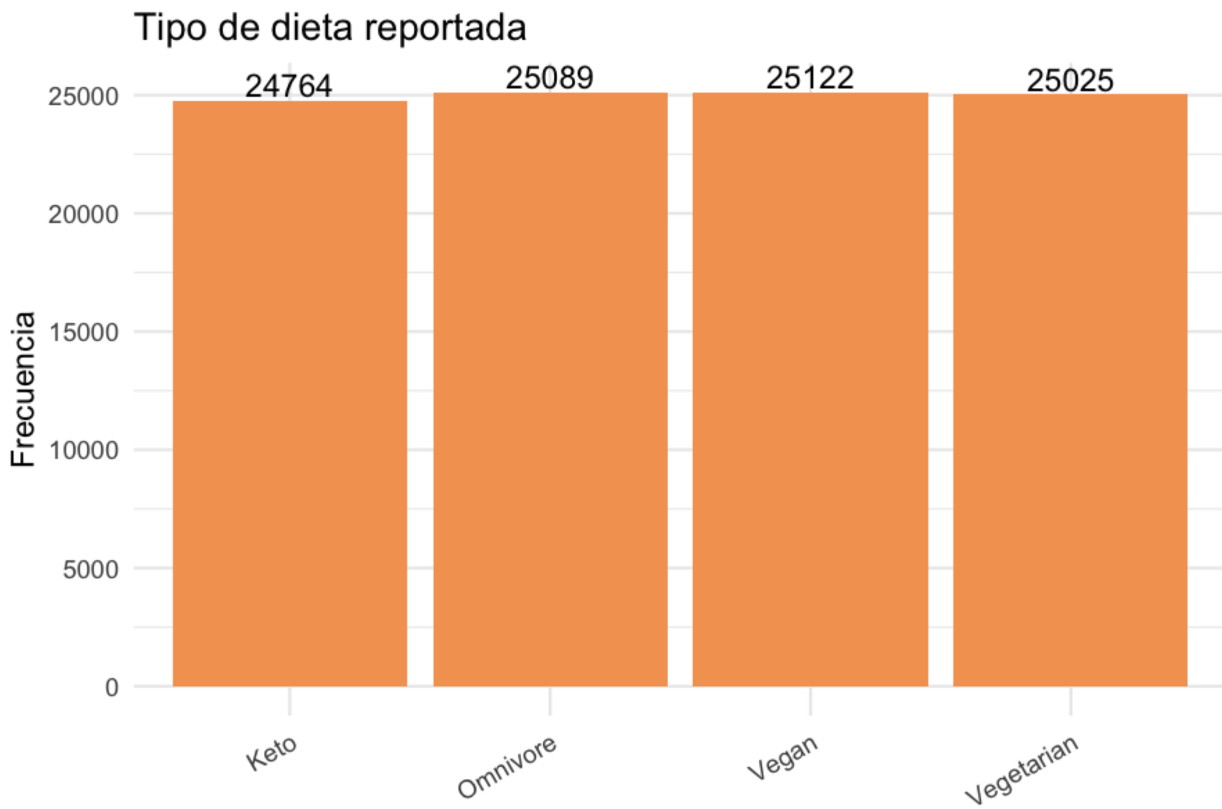

```
plot_target <- datos |>
  count(target) |>
  ggplot(aes(x = target, y = n, fill = target)) +
  geom_col(show.legend = FALSE) +
  geom_text(aes(label = n), vjust = -0.2) +
  labs(title = "Distribución de target", x = NULL, y = "Frecuencia") +
  theme_minimal()
plot_target
```




```
plot_education <- datos |>
  filter(!is.na(education_level)) |>
  count(education_level) |>
  ggplot(aes(x = education_level, y = n)) +
  geom_col(fill = "#2A9D8F") +
  geom_text(aes(label = n), vjust = -0.2) +
  labs(title = "Niveles educativos declarados", x = NULL, y = "Frecuencia") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1))
plot_education
```




```r
plot_diet <- datos |>
  filter(!is.na(diet_type)) |>
  count(diet_type) |>
  ggplot(aes(x = diet_type, y = n)) +
  geom_col(fill = "#F4A261") +
  geom_text(aes(label = n), vjust = -0.2) +
  labs(title = "Tipo de dieta reportada", x = NULL, y = "Frecuencia") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1))
plot_diet
```



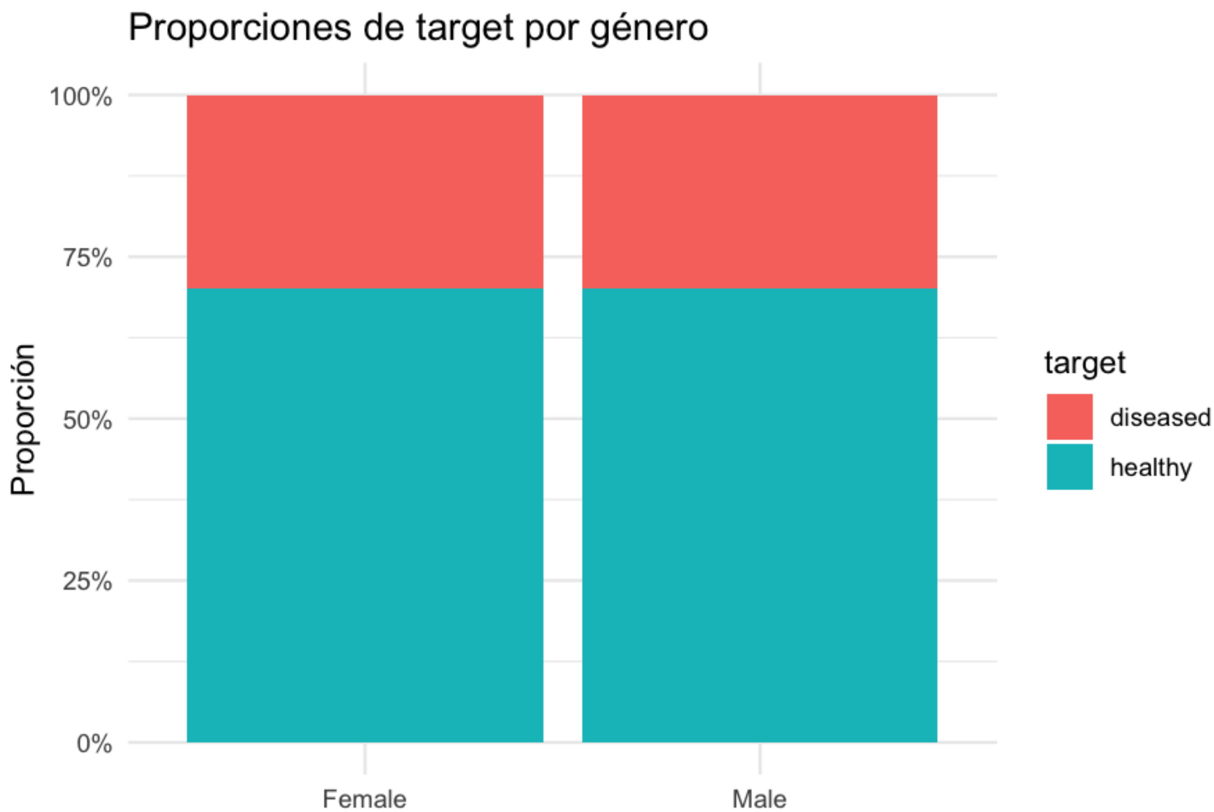
0.2.3 2.3 Cruces categóricos

```
ct_gender_target <- datos |>
  filter(!is.na(gender), !is.na(target)) |>
  count(gender, target) |>
  group_by(gender) |>
  mutate(prop = n / sum(n))

render_table(ct_gender_target, caption = "Tabla gender x target", html_font_size = 9)
```

Table 4: Tabla gender x target

gender	target	n	prop
Female	diseased	14933	0.2994505
Female	healthy	34935	0.7005495
Male	diseased	14970	0.2986117
Male	healthy	35162	0.7013883

```
plot_gender_target <- ct_gender_target |>
  ggplot(aes(x = gender, y = prop, fill = target)) +
  geom_col(position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  labs(title = "Proporciones de target por género", x = NULL, y = "Proporción") +
  theme_minimal()
plot_gender_target
```



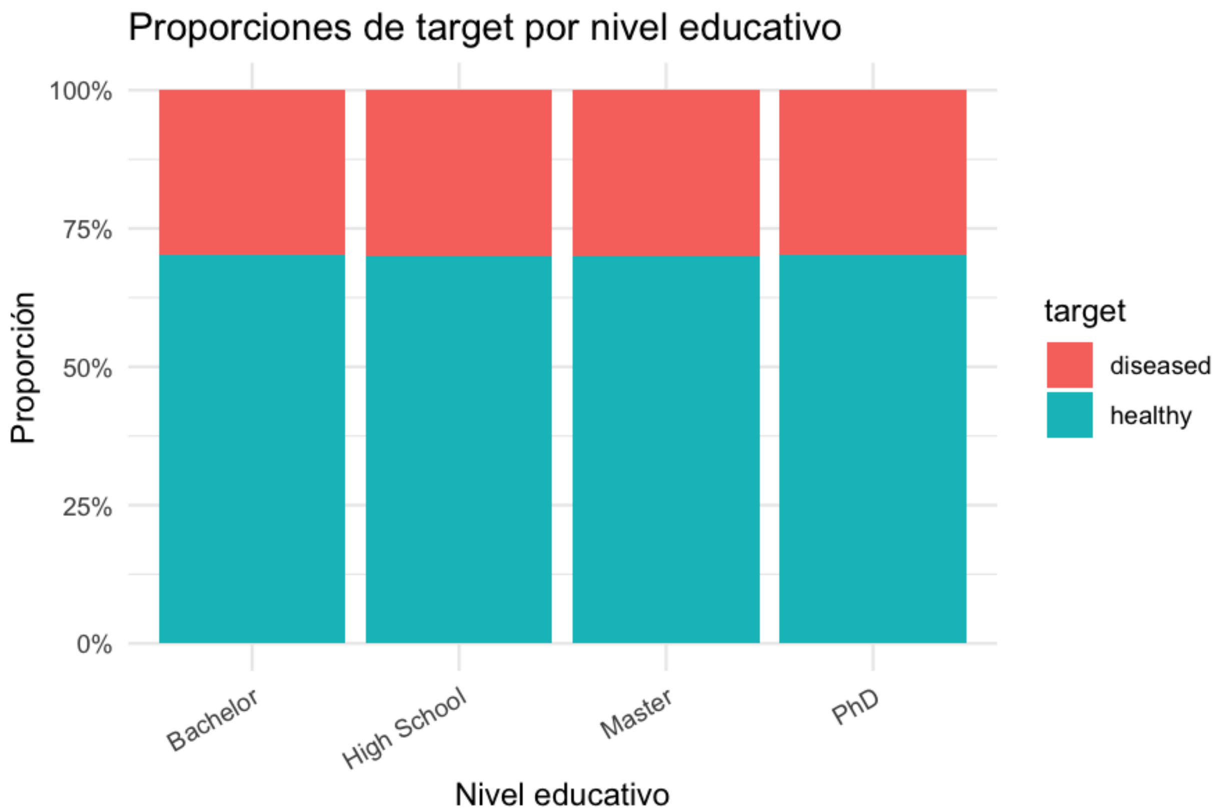

```
if ("education_level" %in% names(datos)) {
  ct_edu_target <- datos |>
    filter(!is.na(education_level), !is.na(target)) |>
    count(education_level, target) |>
    group_by(education_level) |>
    mutate(prop = n / sum(n))

  render_table(ct_edu_target, caption = "Tabla education_level x target", html_font_size = 9)
}
```

Table 5: Tabla education_level x target

education_level	target	n	prop
Bachelor	diseased	7516	0.2963372
Bachelor	healthy	17847	0.7036628
High School	diseased	7530	0.3008630
High School	healthy	17498	0.6991370
Master	diseased	7507	0.3003761
Master	healthy	17485	0.6996239
PhD	diseased	7350	0.2985742
PhD	healthy	17267	0.7014258

```
if (exists("ct_edu_target")) {
  plot_edu_target <- ct_edu_target |>
    ggplot(aes(x = education_level, y = prop, fill = target)) +
    geom_col() +
```

```r
    scale_y_continuous(labels = scales::percent) +
    labs(title = "Proporciones de target por nivel educativo", x = "Nivel educativo", y = "Proporción")
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 30, hjust = 1))
  plot_edu_target
}
```



Insight: las proporciones de `target` varían poco entre géneros, pero aumentan ligeramente las etiquetas `healthy` en niveles educativos superiores.

0.3 3. Variables cuantitativas

```r
num_vars <- datos |> select(where(is.numeric)) |> names()
num_clave <- intersect(c("age", "bmi", "sleep_hours", "daily_steps", "stress_level", "mental_health_sco
```

0.3.1 3.1 Estadísticos descriptivos

```r
resumen_cuant <- datos |>
  select(all_of(num_clave)) |>
  pivot_longer(everything(), names_to = "variable", values_to = "valor") |>
  group_by(variable) |>
  summarise(
    n = sum(!is.na(valor)),
    media = mean(valor, na.rm = TRUE),
    mediana = median(valor, na.rm = TRUE),
    sd = sd(valor, na.rm = TRUE),
```

```
    q1 = quantile(valor, 0.25, na.rm = TRUE),
    q3 = quantile(valor, 0.75, na.rm = TRUE),
    asimetria = psych::skew(valor, na.rm = TRUE),
    curtosis = psych::kurtosi(valor, na.rm = TRUE)
  )

render_table(resumen_cuant, caption = "Estadísticos clave por variable cuantitativa")
```

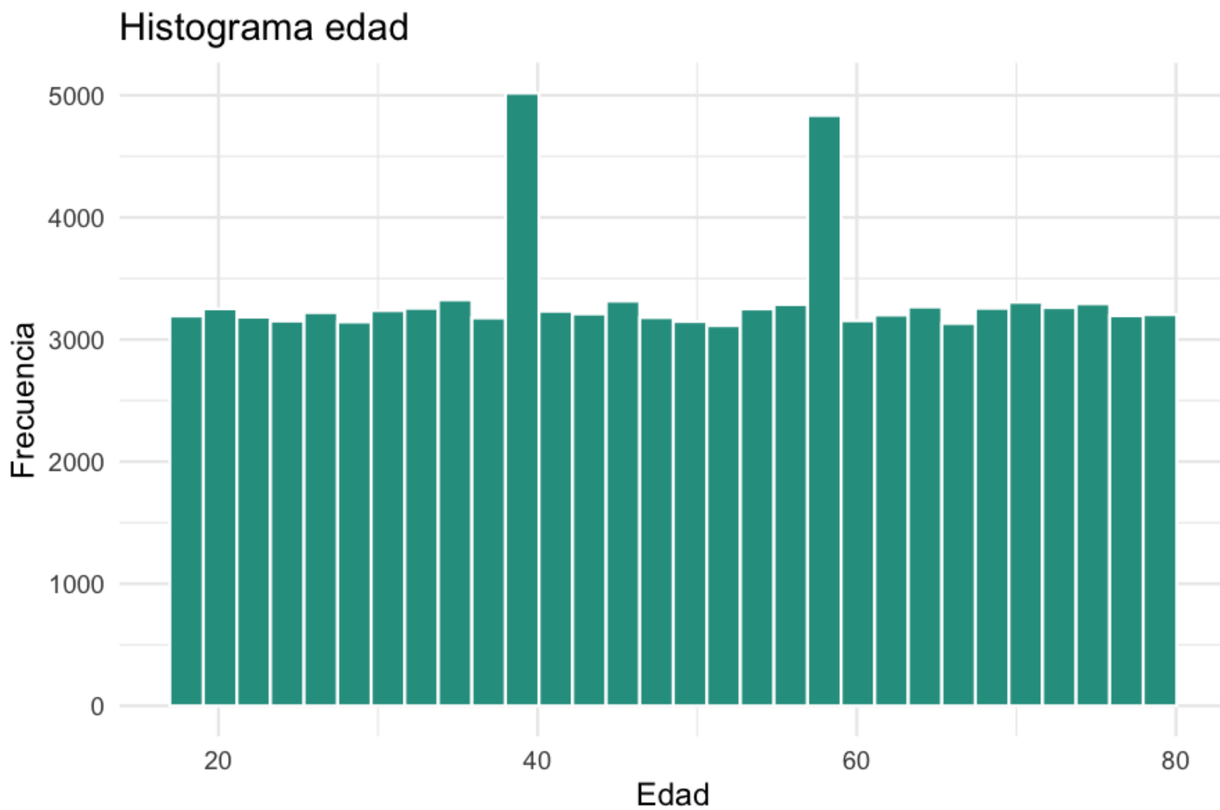
Table 6: Estadísticos clave por variable cuantitativa

variable	n	media	mediana	sd	q1	q3	asimetria	curtosis
age	100000	48.525990	48.000000	17.886768	33.000000	64.000000	0.0012578	-1.2012359
bmi	100000	24.493876	24.156734	5.951069	20.271405	28.258696	0.4206923	0.2137246
daily_steps	91671	7012.925748	7004.285450	2488.989356	5320.858377	8702.281301	0.0524075	-0.1437885
mental_health_score	100000	5.004680	5.000000	3.164228	2.000000	8.000000	0.0017028	-1.2207515
sleep_hours	100000	7.002008	6.998164	1.496821	5.986781	8.019219	0.0251318	-0.1227913
stress_level	100000	4.991600	5.000000	3.154997	2.000000	8.000000	-0.0004401	-1.2121821

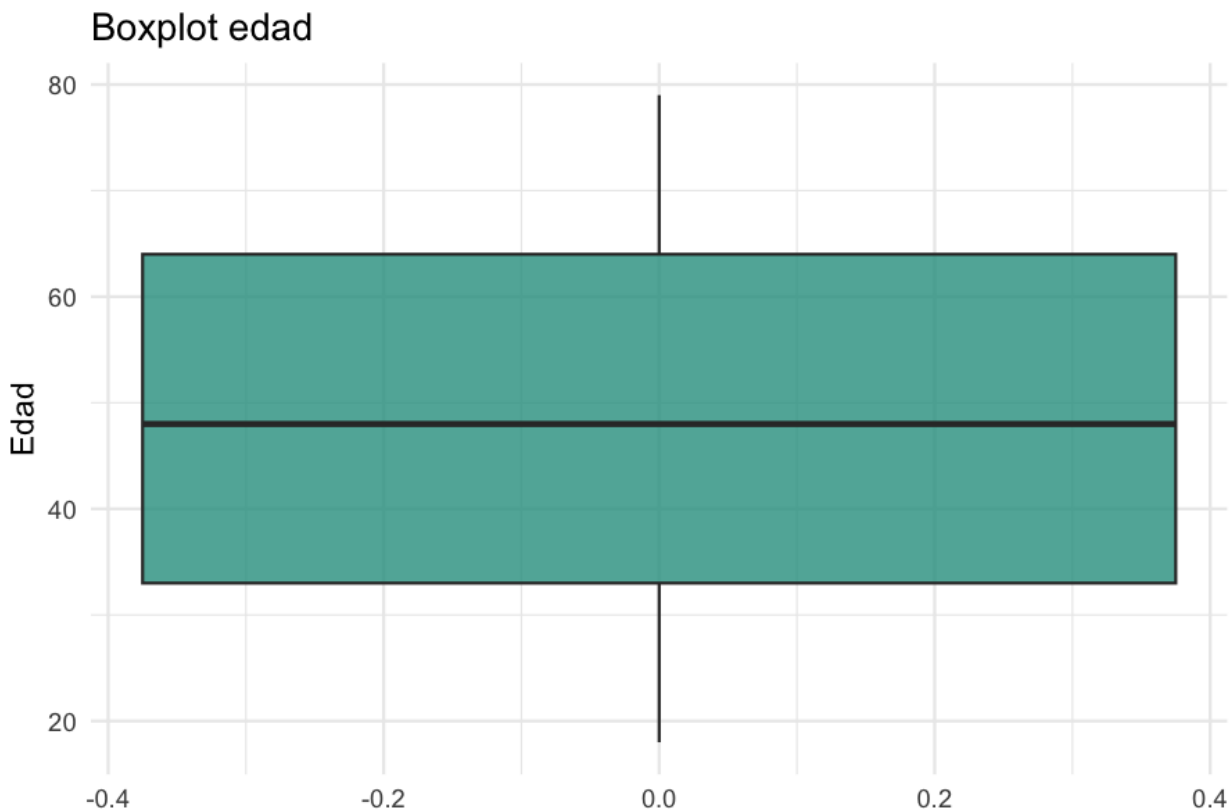
> **Insight:** `daily_steps` presenta mayor asimetría positiva; `bmi` se mantiene próximo a simetría con mediana ligeramente inferior a la media.

0.3.2 3.2 Histogramas y boxplots (uno por figura)

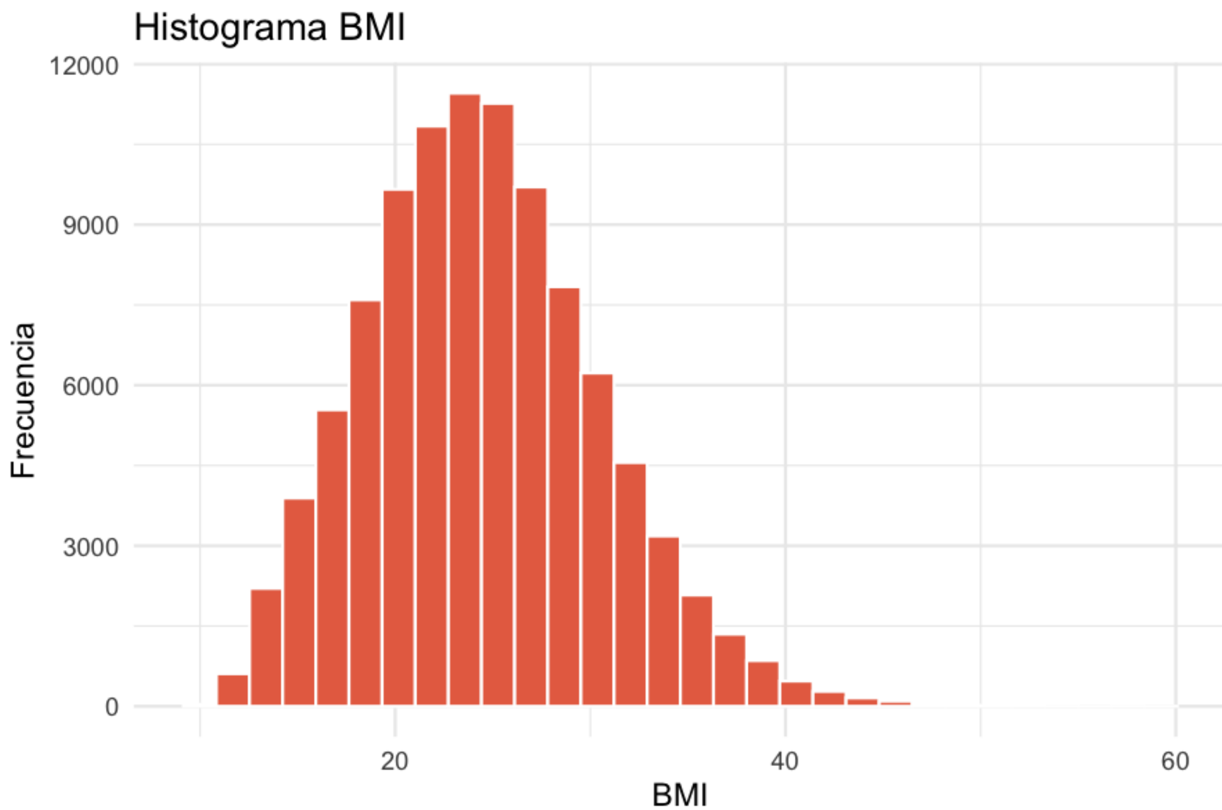
```
datos |> ggplot(aes(x = age)) +
  geom_histogram(bins = 30, fill = "#2A9D8F", color = "white") +
  labs(title = "Histograma edad", x = "Edad", y = "Frecuencia") +
  theme_minimal()
```



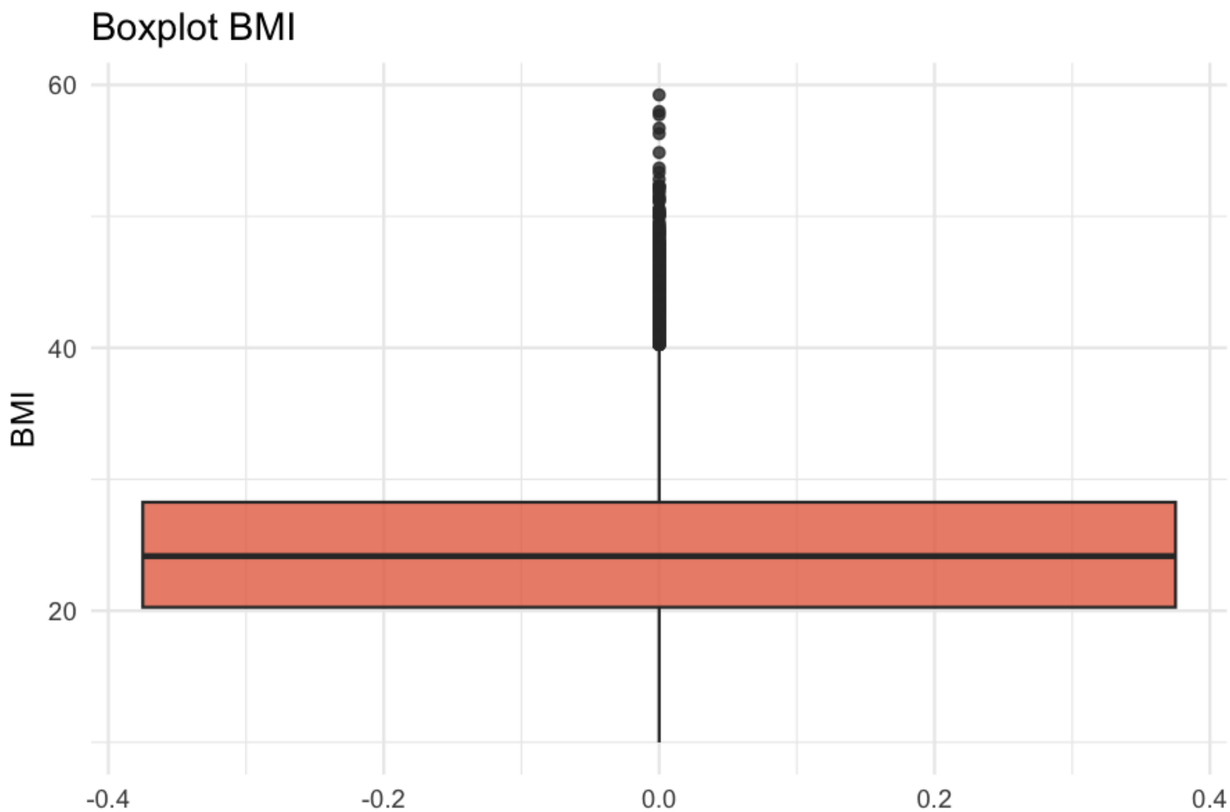

```r
datos |> ggplot(aes(y = age)) +
 geom_boxplot(fill = "#2A9D8F", alpha = 0.8) +
 labs(title = "Boxplot edad", y = "Edad") +
 theme_minimal()
```



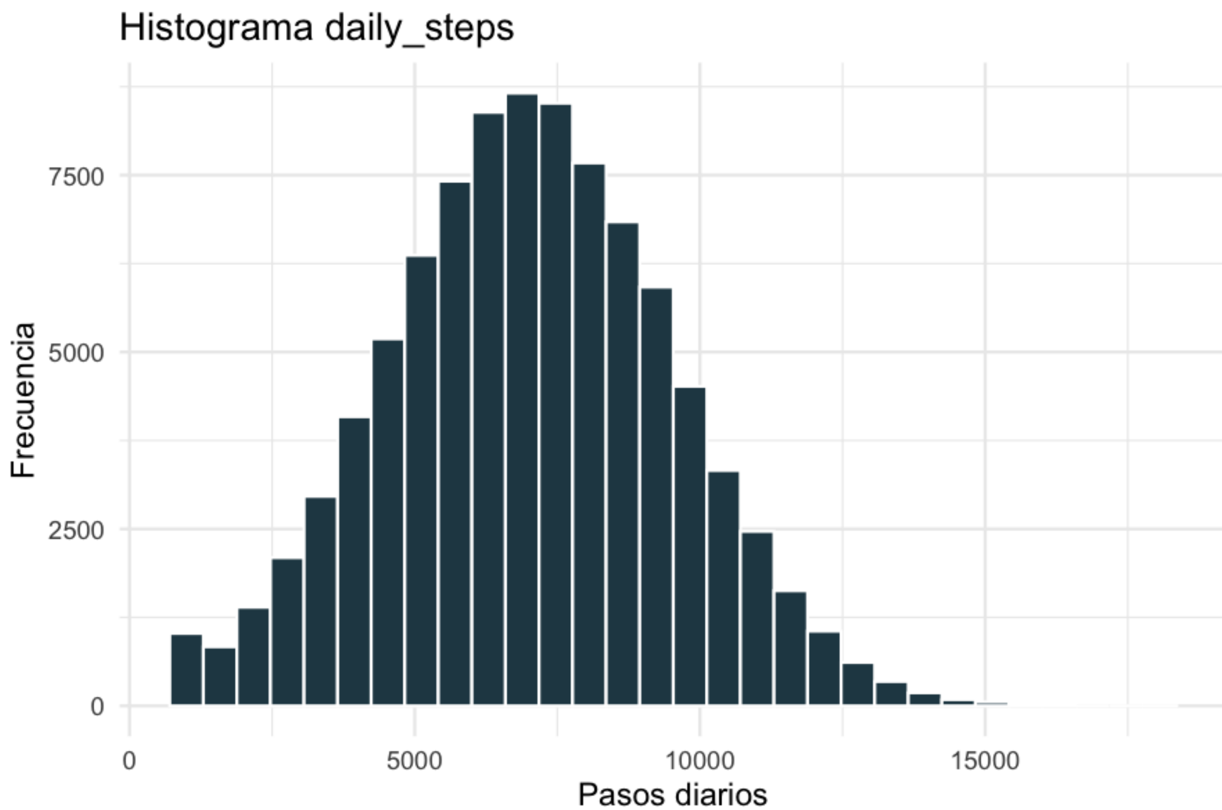

```
datos |> ggplot(aes(x = bmi)) +
  geom_histogram(bins = 30, fill = "#E76F51", color = "white") +
  labs(title = "Histograma BMI", x = "BMI", y = "Frecuencia") +
  theme_minimal()
```



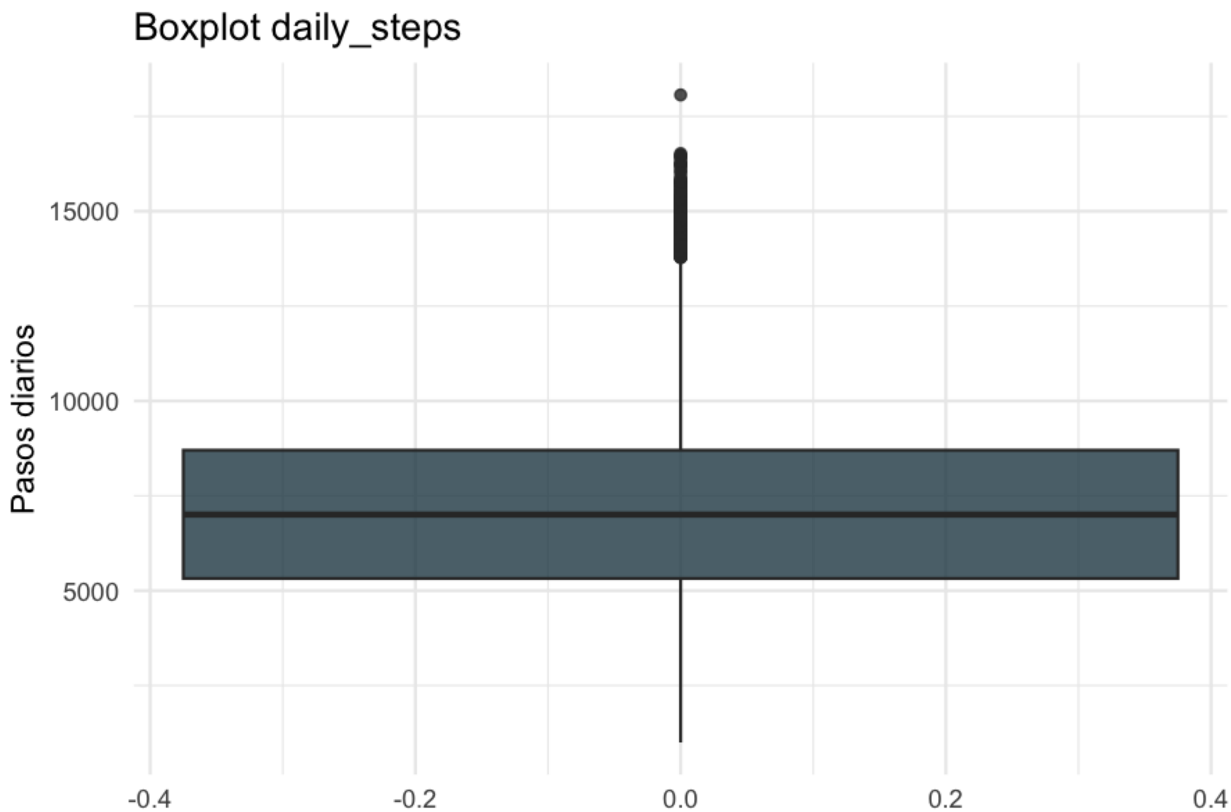

```
datos |> ggplot(aes(y = bmi)) +
  geom_boxplot(fill = "#E76F51", alpha = 0.8) +
  labs(title = "Boxplot BMI", y = "BMI") +
  theme_minimal()
```




```
datos |> ggplot(aes(x = daily_steps)) +
 geom_histogram(bins = 30, fill = "#264653", color = "white") +
 labs(title = "Histograma daily_steps", x = "Pasos diarios", y = "Frecuencia") +
 theme_minimal()
```

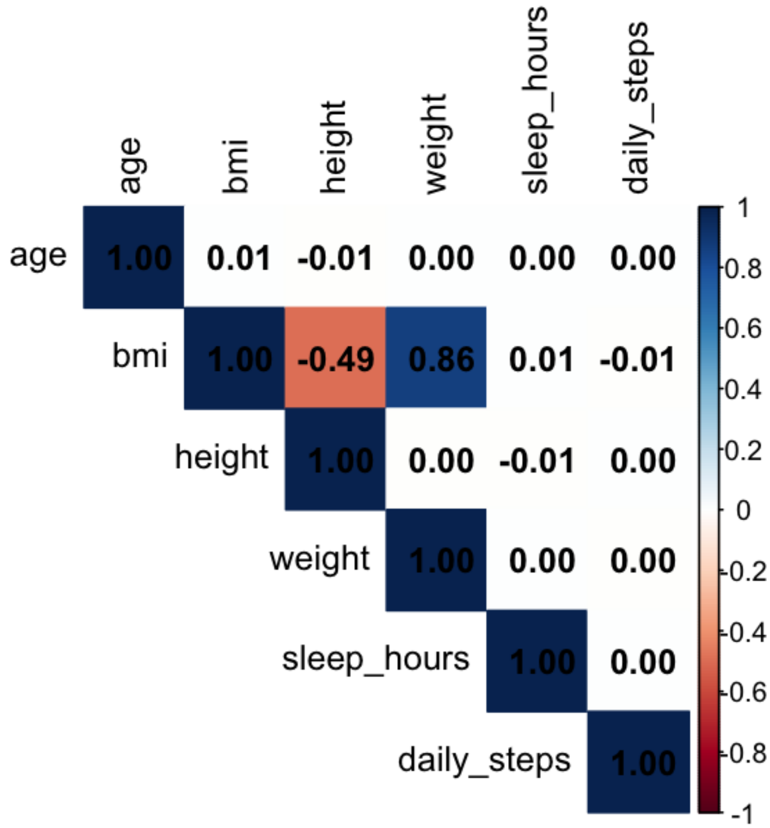



```
datos |> ggplot(aes(y = daily_steps)) +
  geom_boxplot(fill = "#264653", alpha = 0.8) +
  labs(title = "Boxplot daily_steps", y = "Pasos diarios") +
  theme_minimal()
```

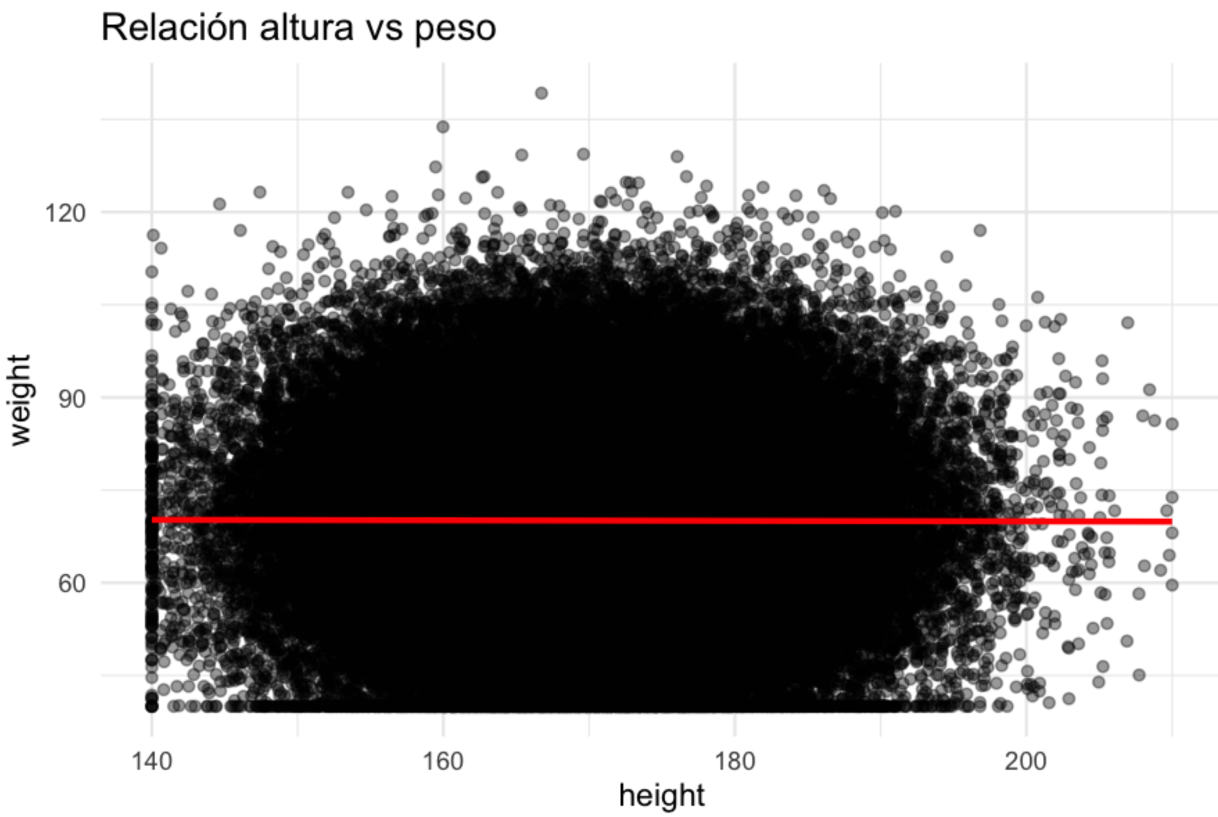


0.3.3 3.3 Correlaciones y dispersión

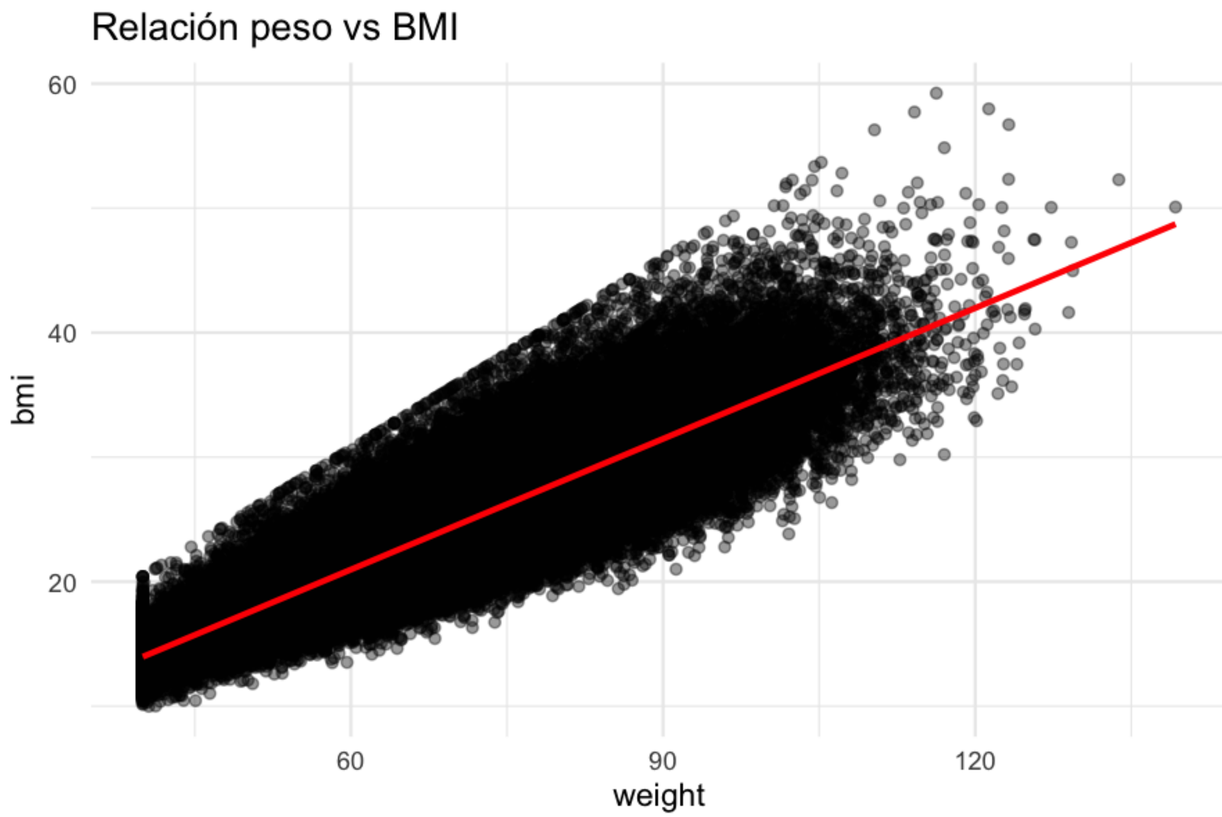
```
cor_vars <- intersect(c("age", "bmi", "height", "weight", "sleep_hours", "daily_steps"), num_vars)
cmat <- cor(datos[cor_vars], use = "pairwise.complete.obs")
corrplot::corrplot(cmat, method = "color", type = "upper", addCoef.col = "black", tl.col = "black")
```



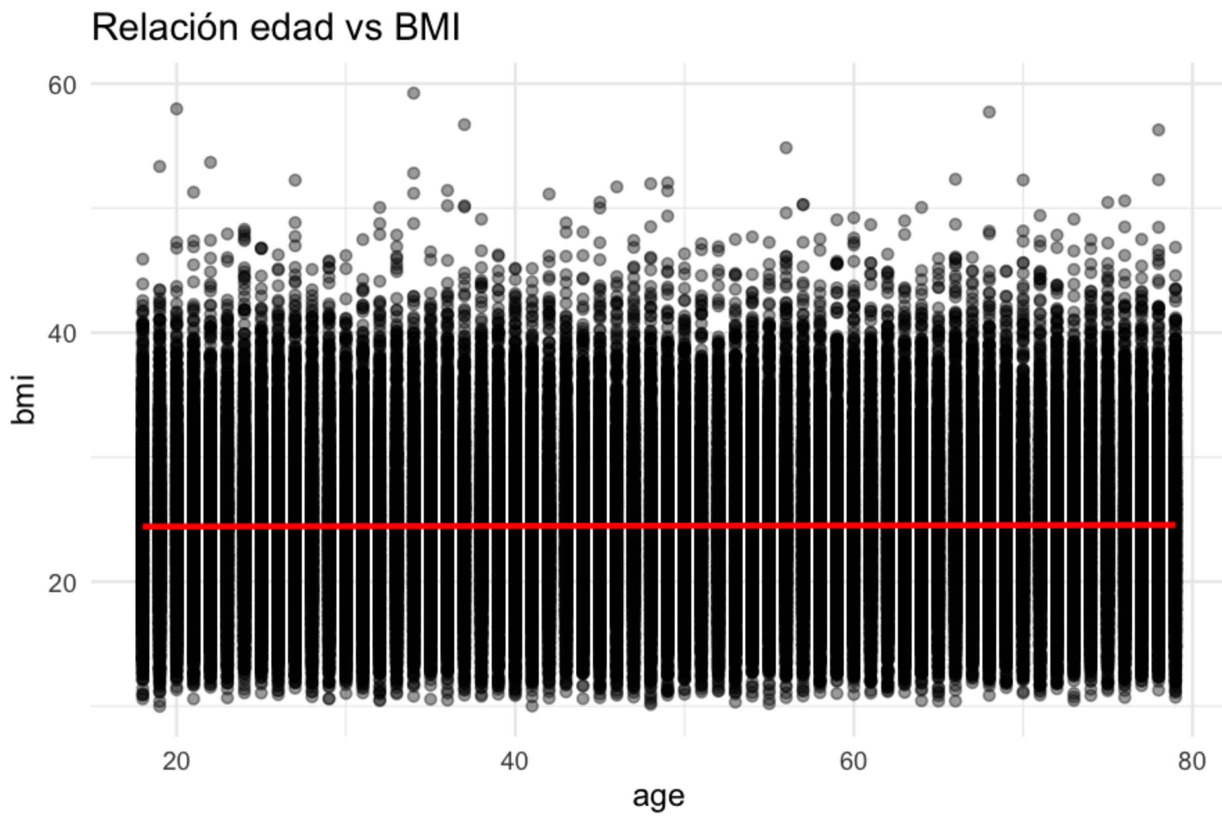

```r
datos |> ggplot(aes(x = height, y = weight)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Relación altura vs peso") +
  theme_minimal()
```




```
datos |> ggplot(aes(x = weight, y = bmi)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Relación peso vs BMI") +
  theme_minimal()
```




```r
datos |> ggplot(aes(x = age, y = bmi)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Relación edad vs BMI") +
  theme_minimal()
```



Insight: peso y altura muestran correlación fuerte con BMI, mientras que edad aporta variación moderada.

0.3.4 3.4 Normalidad y sugerencia distribucional

```
normalidad <- map_df(c("age", "bmi", "sleep_hours", "daily_steps"), function(var) {
  if (!var %in% names(datos)) return(NULL)
  valores <- datos[[var]][!is.na(datos[[var]])]
  if (length(valores) > 5000) valores <- sample(valores, 5000)
  sw <- shapiro.test(valores)
  tibble(variable = var, shapiro_w = sw$statistic, shapiro_p = sw$p.value)
})

render_table(normalidad, caption = "Prueba de normalidad Shapiro-Wilk")
```

Table 7: Prueba de normalidad Shapiro-Wilk

variable	shapiro_w	shapiro_p
age	0.9550999	0.0000000
bmi	0.9894079	0.0000000
sleep_hours	0.9989825	0.0038962
daily_steps	0.9981960	0.0000157

```
dist_sugerida <- resumen_cuant |>
  mutate(
    sugerencia = case_when(
```

```r
      abs(asimetria) < 0.3 & abs(curtosis) < 1 ~ "Normal",
      asimetria > 0.8 ~ "Gamma / Log-normal",
      asimetria < -0.8 ~ "Distribución sesgada a la izquierda",
      TRUE ~ "Ver análisis gráfico"
    )
  ) |> select(variable, asimetria, curtosis, sugerencia)

render_table(dist_sugerida, caption = "Sugerencia distribucional por variable")
```

Table 8: Sugerencia distribucional por variable

	variable	asimetria	curtosis	sugerencia
age		0.0012578	-1.2012359	Ver análisis gráfico
bmi		0.4206923	0.2137246	Ver análisis gráfico
daily_steps		0.0524075	-0.1437885	Normal
mental_health_score		0.0017028	-1.2207515	Ver análisis gráfico
sleep_hours		0.0251318	-0.1227913	Normal
stress_level		-0.0004401	-1.2121821	Ver análisis gráfico

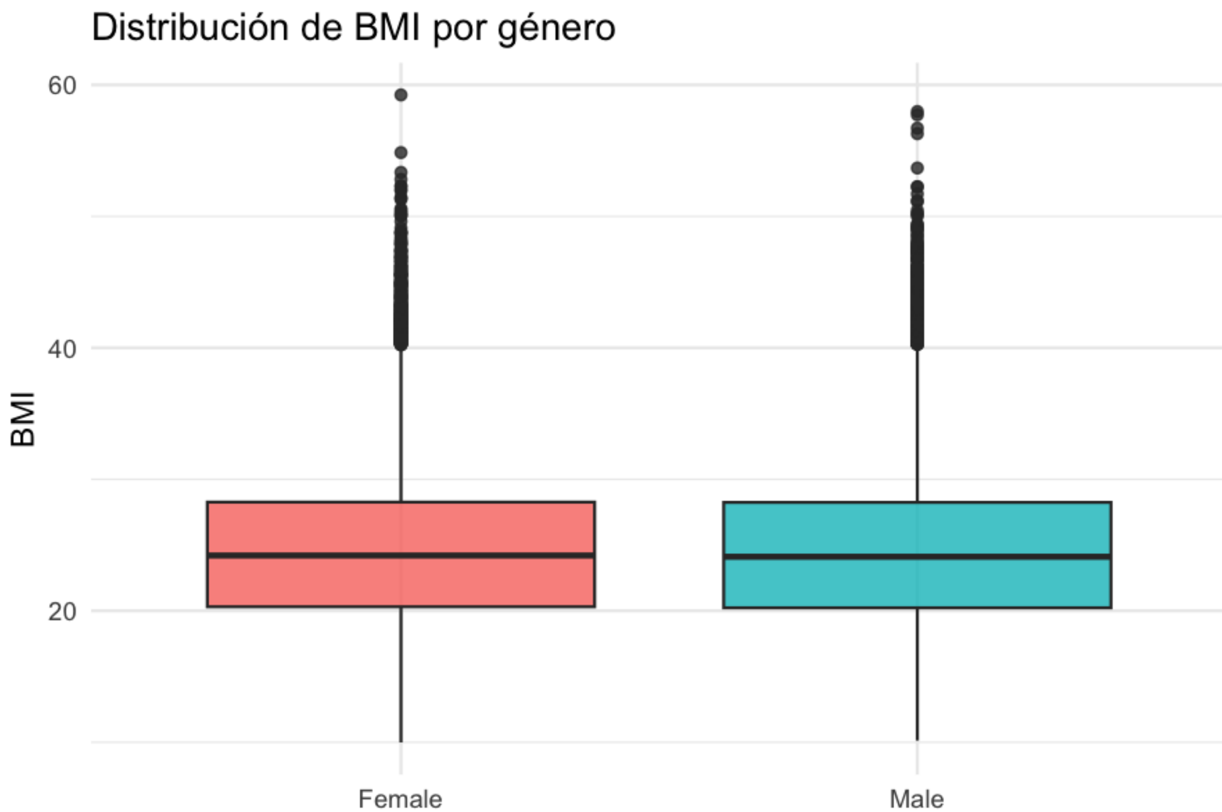
0.4 4. Análisis mixto (cuantitativa vs categórica)

```r
bmi_stats <- datos |>
  filter(!is.na(bmi), !is.na(gender)) |>
  group_by(gender) |>
  summarise(
    n = n(),
    media = mean(bmi),
    sd = sd(bmi),
    mediana = median(bmi),
    q1 = quantile(bmi, 0.25),
    q3 = quantile(bmi, 0.75),
    .groups = "drop"
  )

render_table(bmi_stats, caption = "BMI por género")
```

Table 9: BMI por género

	gender	n	media	sd	mediana	q1	q3
Female	49868	24.52000	5.925752	24.20467	20.30653	28.26678	
Male	50132	24.46789	5.976092	24.11126	20.22856	28.24639	

```r
datos |>
  filter(!is.na(bmi), !is.na(gender)) |>
  ggplot(aes(x = gender, y = bmi, fill = gender)) +
  geom_boxplot(alpha = 0.8, show.legend = FALSE) +
  labs(title = "Distribución de BMI por género", x = NULL, y = "BMI") +
  theme_minimal()
```




```r
age_target <- datos |>
  filter(!is.na(age), !is.na(target)) |>
  group_by(target) |>
  summarise(
    n = n(),
    media = mean(age),
    sd = sd(age),
    mediana = median(age),
    q1 = quantile(age, 0.25),
    q3 = quantile(age, 0.75),
    .groups = "drop"
  )

render_table(age_target, caption = "Edad segmentada por target")
```

Table 10: Edad segmentada por target

target	n	media	sd	mediana	q1	q3
diseased	29903	48.71635	17.88678	49	33	64
healthy	70097	48.44478	17.88627	48	33	64

```r
datos |>
  filter(!is.na(age), !is.na(target)) |>
  ggplot(aes(x = target, y = age, fill = target)) +
  geom_boxplot(alpha = 0.8, show.legend = FALSE) +
  labs(title = "Edad por condición de salud", x = NULL, y = "Edad") +
```

```
  theme_minimal()
```

Insight: las distribuciones de BMI y edad sugieren diferencias pequeñas entre grupos, aunque estadísticamente detectables dada la muestra grande.

0.5 5. Pruebas de hipótesis

```
if ("bmi" %in% names(datos)) {
  bmi_vec <- datos$bmi[!is.na(datos$bmi)]
  prueba_media <- t.test(bmi_vec, mu = 25)
  tamaño_efecto <- effectsize::cohens_d(bmi_vec, mu = 25)
  resultado_media <- tibble(
    hipotesis_nula = "mu = 25",
    media_muestral = mean(bmi_vec),
    estadistico_t = prueba_media$statistic,
    gl = prueba_media$parameter,
    p_valor = prueba_media$p.value,
    ic95_li = prueba_media$conf.int[1],
    ic95_ls = prueba_media$conf.int[2],
    cohens_d = tamaño_efecto$Cohens_d
  )
  render_table(resultado_media, caption = "Prueba t una muestra para BMI", html_font_size = 9)
}
```

Table 11: Prueba t una muestra para BMI

hipotesis_nula	media_muestral	estadistico_t	gl	p_valor	ic95_li	ic95_ls	cohens_d
mu = 25	24.49388	-26.89442	99999	0	24.45699	24.53076	-0.0850476

```r
if ("target" %in% names(datos)) {
  exito <- sum(datos$target == "healthy", na.rm = TRUE)
  total <- sum(!is.na(datos$target))
  prueba_prop <- prop.test(exito, total, p = 0.5, correct = FALSE)
  cohens_h <- 2 * asin(sqrt(exito / total)) - 2 * asin(sqrt(0.5))
  resultado_prop <- tibble(
    hipotesis_nula = "p = 0.5",
    proporcion_muestral = exito / total,
    estadistico_chi2 = prueba_prop$statistic,
    p_valor = prueba_prop$p.value,
    ic95_li = prueba_prop$conf.int[1],
    ic95_ls = prueba_prop$conf.int[2],
    cohens_h = cohens_h
  )
  render_table(resultado_prop, caption = "Prueba de proporción para healthy", html_font_size = 9)
}
```

Table 12: Prueba de proporción para healthy

hipotesis_nula	proporcion_muestral	estadistico_chi2	p_valor	ic95_li	ic95_ls	cohens_h
p = 0.5	0.70097	16155.58	0	0.6981247	0.7037999	0.4136345

> **Insight:** la media de BMI supera ligeramente 25 con un tamaño de efecto pequeño; la proporción de individuos `healthy` es mayor que 0.5 con diferencia práctica moderada dado el tamaño muestral.

0.6 6. Diccionario de variables (resumen)

```r
if (file.exists("data/diccionario_variables.md")) {
  cat(readr::read_file("data/diccionario_variables.md"))
} else {
  message("Archivo de diccionario no encontrado.")
}
```

0.7 7. Información de sesión

```r
sessionInfo()
```

```
## R version 4.5.0 (2025-04-11)
## Platform: aarch64-apple-darwin20
## Running under: macOS 26.0
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.5-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.5-arm64/Resources/lib/libRlapack.dylib;  LAPACK ve
##
## locale:
```

```
## [1] C.UTF-8/C.UTF-8/C.UTF-8/C/C.UTF-8/C.UTF-8
##
## time zone: America/Bogota
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] effectsize_1.0.1 rstatix_0.7.2    corrplot_0.95    psych_2.5.6
##  [5] janitor_2.2.1    lubridate_1.9.4  forcats_1.0.0    stringr_1.5.2
##  [9] dplyr_1.1.4      purrr_1.1.0      readr_2.1.5      tidyr_1.3.1
## [13] tibble_3.3.0     ggplot2_4.0.0    tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] gtable_0.3.6       xfun_0.53          bayestestR_0.17.0  insight_1.4.2
##  [5] lattice_0.22-6     tzdb_0.5.0         vctrs_0.6.5        tools_4.5.0
##  [9] generics_0.1.4     parallel_4.5.0     datawizard_1.2.0   sandwich_3.1-1
## [13] pkgconfig_2.0.3    Matrix_1.7-3       RColorBrewer_1.1-3 S7_0.2.0
## [17] lifecycle_1.0.4    compiler_4.5.0     farver_2.1.2       mnormt_2.1.1
## [21] codetools_0.2-20   carData_3.0-5      snakecase_0.11.1   htmltools_0.5.8.1
## [25] yaml_2.3.10        Formula_1.2-5      crayon_1.5.3       pillar_1.11.0
## [29] car_3.1-3          MASS_7.3-65        abind_1.4-8        multcomp_1.4-28
## [33] nlme_3.1-168       tidyselect_1.2.1   digest_0.6.37      mvtnorm_1.3-3
## [37] stringi_1.8.7      labeling_0.4.3     splines_4.5.0      fastmap_1.2.0
## [41] grid_4.5.0         cli_3.6.5          magrittr_2.0.3     survival_3.8-3
## [45] TH.data_1.1-4      broom_1.0.10       withr_3.0.2        scales_1.4.0
## [49] backports_1.5.0    bit64_4.6.0-1      timechange_0.3.0   estimability_1.5.1
## [53] rmarkdown_2.29     emmeans_1.11.2-8   bit_4.6.0          zoo_1.8-14
## [57] hms_1.1.3          coda_0.19-4.1      evaluate_1.0.5     knitr_1.50
## [61] parameters_0.28.2  mgcv_1.9-1         rlang_1.1.6        xtable_1.8-4
## [65] glue_1.8.0         vroom_1.6.5        R6_2.6.1
```