Entrega 1 - Notebook Reproducible

[Integrantes del Grupo]

2025-09-17

Contents

1. Carga de Librerías

```
suppressPackageStartupMessages({
  library(tidyverse)
  library(janitor)
  library(psych)
  library(readr)
  library(corrplot)
  library(rstatix)
  library(effectsize)
  library(patchwork)
})
render_table <- function(data, caption = NULL, html_font_size = 10, ...) {
```

```
  tab <- knitr::kable(data, caption = caption, ...)
  if (knitr::is_html_output()) {
    tab <- kableExtra::kable_styling(tab, font_size = html_font_size)
  }
  tab
}
```

2. Carga y Exploración Inicial de Datos

```
# Carga reproducible (ruta relativa)
df <- read_csv("/Users/thom/Library/Mobile Documents/com~apple~CloudDocs/Universidad/Proba 2/Probability
cat("Filas:", nrow(df), " - Columnas:", ncol(df), "\n")
```

```
## Filas: 100000  - Columnas: 48
```

```
str(df, give.attr = FALSE)
```

```
## spc_tbl_ [100,000 x 48] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ survey_code             : num [1:100000] 1 2 3 4 5 6 7 8 9 10 ...
##  $ age                     : num [1:100000] 56 69 46 32 60 25 78 38 56 75 ...
##  $ gender                  : chr [1:100000] "Male" "Female" "Male" "Female" ...
##  $ height                  : num [1:100000] 173 163 177 172 164 ...
##  $ weight                  : num [1:100000] 56.9 97.8 80.7 63.1 40 ...
##  $ bmi                     : num [1:100000] 18.9 36.7 25.7 21.3 14.9 ...
##  $ bmi_estimated           : num [1:100000] 18.9 36.7 25.7 21.3 14.9 ...
##  $ bmi_scaled              : num [1:100000] 56.7 110.1 77 64 44.8 ...
##  $ bmi_corrected           : num [1:100000] 19 36.5 25.6 21.2 14.8 ...
##  $ waist_size              : num [1:100000] 72.2 85.6 90.3 100.5 69 ...
##  $ blood_pressure          : num [1:100000] 118 118 123 148 151 ...
##  $ heart_rate              : num [1:100000] 60.7 66.5 76 68.8 92.3 ...
##  $ cholesterol             : num [1:100000] 215 116 138 203 200 ...
##  $ glucose                 : num [1:100000] 103 116.9 89.2 128.4 94.8 ...
##  $ insulin                 : num [1:100000] NA 10.1 NA 18.7 16 ...
##  $ sleep_hours             : num [1:100000] 6.48 8.43 5.7 5.19 7.91 ...
##  $ sleep_quality           : chr [1:100000] "Fair" "Good" "Poor" "Good" ...
##  $ work_hours              : num [1:100000] 7.67 9.52 5.83 9.49 7.28 ...
##  $ physical_activity       : num [1:100000] 0.357 0.568 3.764 0.889 2.902 ...
##  $ daily_steps             : num [1:100000] 13321 11911 2974 5322 9791 ...
##  $ calorie_intake          : num [1:100000] 2674 2650 1747 2034 2386 ...
##  $ sugar_intake            : num [1:100000] 44.5 74.7 19.7 82.6 46 ...
##  $ alcohol_consumption     : chr [1:100000] NA "Regularly" "Regularly" "Occasionally" ...
##  $ smoking_level           : chr [1:100000] "Non-smoker" "Light" "Heavy" "Heavy" ...
##  $ water_intake            : num [1:100000] 1.694 0.716 2.488 2.643 1.968 ...
##  $ screen_time             : num [1:100000] 5 5.93 4.37 4.12 3.18 ...
##  $ stress_level            : num [1:100000] 2 3 0 10 9 7 7 7 2 10 ...
##  $ mental_health_score     : num [1:100000] 8 9 1 4 7 6 1 2 9 9 ...
##  $ mental_health_support   : chr [1:100000] "No" "No" "No" "No" ...
##  $ education_level         : chr [1:100000] "PhD" "High School" "Master" "Master" ...
##  $ job_type                : chr [1:100000] "Tech" "Office" "Office" "Labor" ...
##  $ occupation              : chr [1:100000] "Farmer" "Engineer" "Teacher" "Teacher" ...
##  $ income                  : num [1:100000] 6760 6241 3429 2619 3662 ...
##  $ diet_type               : chr [1:100000] "Vegan" "Vegan" "Vegan" "Vegetarian" ...
##  $ exercise_type           : chr [1:100000] "Strength" "Cardio" "Cardio" "Mixed" ...
```

```
##  $ device_usage            : chr [1:100000] "High" "Moderate" "High" "Low" ...
##  $ healthcare_access       : chr [1:100000] "Poor" "Moderate" "Good" "Moderate" ...
##  $ insurance               : chr [1:100000] "No" "No" "Yes" "No" ...
##  $ sunlight_exposure       : chr [1:100000] "High" "High" "High" "High" ...
##  $ meals_per_day           : num [1:100000] 5 5 4 1 1 4 2 3 2 1 ...
##  $ caffeine_intake         : chr [1:100000] "Moderate" "High" "Moderate" "None" ...
##  $ family_history          : chr [1:100000] "No" "Yes" "No" "No" ...
##  $ pet_owner               : chr [1:100000] "Yes" "No" "No" "Yes" ...
##  $ electrolyte_level       : num [1:100000] 0 0 0 0 0 0 0 0 0 0 ...
##  $ gene_marker_flag        : num [1:100000] 1 1 1 1 1 1 NA 1 NA 1 ...
##  $ environmental_risk_score: num [1:100000] 5.5 5.5 5.5 5.5 5.5 5.5 5.5 5.5 5.5 5.5 ...
##  $ daily_supplement_dosage : num [1:100000] -2.276 6.239 5.424 8.389 0.333 ...
##  $ target                  : chr [1:100000] "healthy" "healthy" "healthy" "healthy" ...
```

```
missing_tbl <- df %>% summarise(across(everything(), ~sum(is.na(.)))) %>% pivot_longer(everything(), nam
missing_tbl %>%
  head(15) %>%
  render_table(col.names = c("Variable","NA","%NA"), caption = "Top 15 variables con mayor NA", html_for
```

Table 1: Top 15 variables con mayor NA

Variable	NA	%NA
insulin	15836	15.84
heart_rate	14003	14.00
alcohol_consumption	13910	13.91
gene_marker_flag	10474	10.47
income	8470	8.47
daily_steps	8329	8.33
blood_pressure	7669	7.67
survey_code	0	0.00
age	0	0.00
gender	0	0.00
height	0	0.00
weight	0	0.00
bmi	0	0.00
bmi_estimated	0	0.00
bmi_scaled	0	0.00

3. Tipología de Variables

```
chr_vars <- names(df %>% select(where(is.character)))
num_vars <- names(df %>% select(where(is.numeric)))
list(chr_vars = chr_vars[1:20], num_vars = num_vars[1:20])
```

```
## $chr_vars
##  [1] "gender"                "sleep_quality"         "alcohol_consumption"
##  [4] "smoking_level"         "mental_health_support" "education_level"
##  [7] "job_type"              "occupation"            "diet_type"
## [10] "exercise_type"         "device_usage"          "healthcare_access"
## [13] "insurance"             "sunlight_exposure"     "caffeine_intake"
## [16] "family_history"        "pet_owner"             "target"
## [19] NA                      NA
##
```

```
## $num_vars
##  [1] "survey_code"       "age"               "height"
##  [4] "weight"            "bmi"               "bmi_estimated"
##  [7] "bmi_scaled"        "bmi_corrected"     "waist_size"
## [10] "blood_pressure"    "heart_rate"        "cholesterol"
## [13] "glucose"           "insulin"           "sleep_hours"
## [16] "work_hours"        "physical_activity" "daily_steps"
## [19] "calorie_intake"    "sugar_intake"
```

4. Marco Metodológico Completo

Origen de los Datos

Los datos provienen del dataset "Health Lifestyle Classification" que contiene información recolectada sobre hábitos de salud y estilo de vida. El dataset incluye variables demográficas, biométricas, de comportamiento, salud mental y socioeconómicas recopiladas mediante encuestas estructuradas.

Objetivo del Análisis

Caracterizar los patrones de salud y estilo de vida en la población estudiada mediante técnicas de estadística descriptiva e inferencial, específicamente: 1. Describir distribuciones univariadas y bivariadas 2. Identificar relaciones entre variables 3. Evaluar si el BMI promedio poblacional difiere de 25 kg/m² 4. Determinar si la proporción de individuos saludables difiere del 50%

5. Análisis Descriptivo de Variables Categóricas

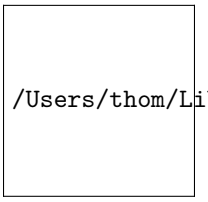
```
cat_vars <- chr_vars
cat_summary <- map_df(cat_vars, ~df %>% count(.data[[.x]]) %>% mutate(variable = .x) %>% rename(valor =
cat_summary %>%
  group_by(variable) %>%
  mutate(pct = round(100*frecuencia/sum(frecuencia),2)) %>%
  ungroup() %>%
  head(40) %>%
  render_table(caption = "Frecuencias (primeras 40 filas)", html_font_size = 10)
```

Table 2: Frecuencias (primeras 40 filas)

valor	frecuencia	variable	pct
Female	49868	gender	49.87
Male	50132	gender	50.13
Excellent	25091	sleep_quality	25.09
Fair	25008	sleep_quality	25.01
Good	25147	sleep_quality	25.15
Poor	24754	sleep_quality	24.75
None	28477	alcohol_consumption	28.48
Occasionally	28831	alcohol_consumption	28.83
Regularly	28782	alcohol_consumption	28.78
NA	13910	alcohol_consumption	13.91
Heavy	33208	smoking_level	33.21
Light	33437	smoking_level	33.44
Non-smoker	33355	smoking_level	33.35
No	50104	mental_health_support	50.10

valor	frecuencia	variable	pct
Yes	49896	mental_health_support	49.90
Bachelor	25363	education_level	25.36
High School	25028	education_level	25.03
Master	24992	education_level	24.99
PhD	24617	education_level	24.62
Healthcare	16546	job_type	16.55
Labor	16777	job_type	16.78
Office	16704	job_type	16.70
Service	16571	job_type	16.57
Tech	16691	job_type	16.69
Unemployed	16711	job_type	16.71
Artist	16657	occupation	16.66
Doctor	16927	occupation	16.93
Driver	16562	occupation	16.56
Engineer	16474	occupation	16.47
Farmer	16719	occupation	16.72
Teacher	16661	occupation	16.66
Keto	24764	diet_type	24.76
Omnivore	25089	diet_type	25.09
Vegan	25122	diet_type	25.12
Vegetarian	25025	diet_type	25.02
Cardio	24988	exercise_type	24.99
Mixed	24778	exercise_type	24.78
None	24969	exercise_type	24.97
Strength	25265	exercise_type	25.26
High	33562	device_usage	33.56

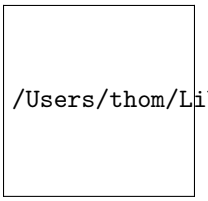
```
if(!requireNamespace("treemapify", quietly = TRUE)) {
  # opcional: instalar en entorno local si falta
  # install.packages("treemapify")
}
if("occupation" %in% names(df) && require(treemapify)) {
  df %>%
    count(occupation) %>%
    filter(!is.na(occupation)) %>%
    ggplot(aes(area = n, fill = n, label = occupation)) +
    geom_treemap() +
    geom_treemap_text(reflow = TRUE, colour = "white", place = "centre", grow = TRUE) +
    labs(title = "Treemap de Occupation", fill = "Frecuencia") +
    theme_minimal()
}
```



/Users/thom/Library/Mobile Documents/com~apple~CloudDocs/Universidad/Proba 2/Probability-Proyect/report

```
key_cat <- intersect(c("gender","target","education_level","diet_type"), names(df))
plot_list <- map(key_cat, ~{
  ggplot(df, aes(x = .data[[.x]])) +
```

```
    geom_bar(fill = "steelblue", alpha = 0.8) +
    geom_text(stat = "count", aes(label = after_stat(count)), vjust = -0.3, size = 3) +
    labs(title = .x, x = NULL, y = "Frecuencia") +
    theme_minimal()
})
wrap_plots(plotlist = plot_list)
```



/Users/thom/Library/Mobile Documents/com~apple~CloudDocs/Universidad/Proba 2/Probability-Proyect/report

```
if(all(c("gender","target") %in% names(df))) {
  tab_gt <- df %>% filter(!is.na(gender), !is.na(target)) %>% count(gender, target) %>% group_by(gender
  tab_gt %>% render_table(caption = "Tabla de contingencia: gender x target")
  # Chi-cuadrado y Cramer's V
  chi_tab <- table(df$gender, df$target)
  chi_obj <- suppressWarnings(chisq.test(chi_tab))
  cramer <- effectsize::cramers_v(chi_tab)
  cat("\nChi-cuadrado:", round(chi_obj$statistic,3), "(gl=", chi_obj$parameter, ") p=", signif(chi_obj$p
  cat("Cramer's V:", round(cramer$Cramers_v,3), "Magnitud:", cramer$Magnitude, "\n")
}
```

```
##
## Chi-cuadrado: 0.08 (gl= 1 ) p= 0.777
## Cramer's V: 0 Magnitud:
```

```
# Gráfico de barras apiladas
if(all(c("gender","target") %in% names(df))) {
  p_stacked <- df %>% filter(!is.na(gender), !is.na(target)) %>%
    ggplot(aes(x = gender, fill = target)) +
    geom_bar(position = "fill") +
    geom_text(stat = "count", aes(label = after_stat(count)), position = position_fill(vjust = 0.5)) +
    labs(title = "Proporción de Target por Género", x = "Género", y = "Proporción", fill = "Target") +
    theme_minimal()

  # Si hay education_level, hacer otro bivariado
  if("education_level" %in% names(df)) {
    p_education <- df %>% filter(!is.na(education_level), !is.na(target)) %>%
      ggplot(aes(x = education_level, fill = target)) +
      geom_bar(position = "dodge") +
      labs(title = "Target por Nivel Educativo", x = "Educación", y = "Frecuencia", fill = "Target") +
      theme_minimal() +
      theme(axis.text.x = element_text(angle = 45, hjust = 1))

    p_stacked + p_education
  } else {
    p_stacked
  }
}
```

/Users/thom/Library/Mobile Documents/com~apple~CloudDocs/Universidad/Proba 2/Probability-Proyect/report

6. Análisis Descriptivo de Variables Cuantitativas

```
num_vars <- setdiff(num_vars, c("survey_code"))
desc_tbl <- df %>% select(any_of(num_vars)) %>% psych::describe() %>% as.data.frame() %>% rownames_to_c
desc_tbl %>%
  filter(variable %in% c("age","bmi","height","weight","sleep_hours","daily_steps")) %>%
  render_table(caption = "Estadígrafos variables clave (básicos)", html_font_size = 9)
```

Table 3: Estadígrafos variables clave (básicos)

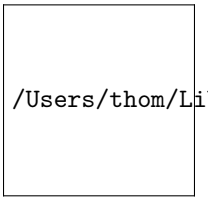
variable	n	mean	sd	median	min	max	skew	kurtosis
age	100000	48.525990	17.886768	48.000000	18.000000	79.00000	0.0012578	-1.2012359
height	100000	170.023707	9.982798	170.016778	140.000000	210.00000	0.0089420	-0.0269878
weight	100000	70.064862	14.693667	69.924141	40.000000	139.25089	0.1288708	-0.2507124
bmi	100000	24.493876	5.951069	24.156734	9.988494	59.23479	0.4206923	0.2137246
sleep_hours	100000	7.002008	1.496821	6.998164	3.000000	12.00000	0.0251318	-0.1227913
daily_steps	91671	7012.925748	2488.989356	7004.285450	1000.000000	18064.96954	0.0524075	-0.1437885

```
mode_vec <- function(x){ux <- na.omit(unique(x)); ux[which.max(tabulate(match(x, ux)))]}
num_extended <- df %>% select(any_of(num_vars)) %>% pivot_longer(everything(), names_to="variable", valu
  n = sum(!is.na(value)),
  mean = mean(value, na.rm=TRUE),
  median = median(value, na.rm=TRUE),
  mode = mode_vec(value),
  sd = sd(value, na.rm=TRUE),
  iqr = IQR(value, na.rm=TRUE),
  p10 = quantile(value,0.10, na.rm=TRUE),
  p25 = quantile(value,0.25, na.rm=TRUE),
  p75 = quantile(value,0.75, na.rm=TRUE),
  p90 = quantile(value,0.90, na.rm=TRUE),
  skew = psych::skew(value, na.rm=TRUE),
  kurt = psych::kurtosi(value, na.rm=TRUE)
)
num_extended %>%
  filter(variable %in% c("age","bmi","sleep_hours","daily_steps")) %>%
  render_table(caption="Estadísticos extendidos (forma y posición)", html_font_size = 9)
```

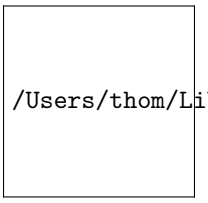
Table 4: Estadísticos extendidos (forma y posición)

variable	n	mean	median	mode	sd	iqr	p10	p25	p75	p90	skew	kurt
age	100000	48.525994	48.000000	71.000000	17.886768	31.000000	24.000000	33.000000	64.000000	73.000000	0.001257	-1.201236
bmi	100000	24.493876	24.156734	20.408165	5.951069	7.987291	17.000839	20.271405	28.258696	32.316226	0.420692	-0.321372
daily_steps	91671	7012.925748	4284.285450	500.000000	7088.989338	8381.422273	735.179482	2020.858377	10272.281301	15215.523575	1.575240	1.437885
sleep_hours	100000	7.002008	6.998164	3.00000	1.496821	2.032438	5.071756	5.986781	8.019219	8.922488	0.025131	-0.122791

```
vars_plot <- intersect(c("age","bmi","sleep_hours","daily_steps"), names(df))
plots_h <- map(vars_plot, ~{
  ggplot(df, aes(x = .data[[.x]])) +
    geom_histogram(bins = 30, fill = "#69b3a2", color = "white") +
    labs(title = paste("Histograma", .x), x = .x, y = "freq") +
    theme_minimal()
})
wrap_plots(plotlist = plots_h)
```

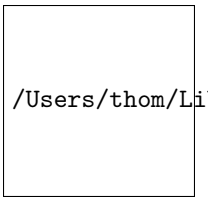

/Users/thom/Library/Mobile Documents/com~apple~CloudDocs/Universidad/Proba 2/Probability-Proyect/report

```
plots_b <- map(vars_plot, ~{
  ggplot(df, aes(y = .data[[.x]])) +
    geom_boxplot(fill = "#ffab00", alpha = 0.7) +
    labs(title = paste("Boxplot", .x), y = .x) +
    theme_minimal()
})
wrap_plots(plotlist = plots_b, ncol = length(plots_b))
```

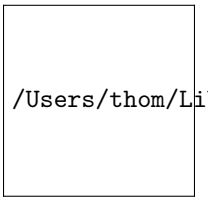

/Users/thom/Library/Mobile Documents/com~apple~CloudDocs/Universidad/Proba 2/Probability-Proyect/report

7. Correlaciones

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

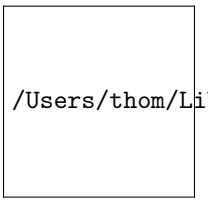



```r
# Diagramas de dispersión para correlaciones altas
if(all(c("height","weight","bmi","age") %in% names(df))) {
  p1 <- ggplot(df, aes(x = height, y = weight)) + geom_point(alpha = 0.6) + geom_smooth(method = "lm", s
  p2 <- ggplot(df, aes(x = weight, y = bmi)) + geom_point(alpha = 0.6) + geom_smooth(method = "lm", se =
  p3 <- ggplot(df, aes(x = height, y = bmi)) + geom_point(alpha = 0.6) + geom_smooth(method = "lm", se =
  p4 <- ggplot(df, aes(x = age, y = bmi)) + geom_point(alpha = 0.6) + geom_smooth(method = "lm", se = F
  (p1 + p2) / (p3 + p4)
}
```



6.1 Análisis Distribucional

```r
# QQ-plots para evaluar normalidad
key_vars <- intersect(c("age","bmi","sleep_hours","daily_steps"), names(df))
qq_plots <- map(key_vars, ~{
  ggplot(df, aes(sample = .data[[.x]])) +
    stat_qq() + stat_qq_line() +
    labs(title = paste("QQ-plot:", .x)) +
    theme_minimal()
})
wrap_plots(plotlist = qq_plots)
```




```r
# Tests de normalidad
normality_results <- map_df(key_vars, ~{
  if(.x %in% names(df)) {
    vec <- df %>% filter(!is.na(.data[[.x]])) %>% pull(.data[[.x]])
    if(length(vec) > 5000) vec <- sample(vec, 5000)
    sw_test <- shapiro.test(vec)
    data.frame(variable = .x, shapiro_w = sw_test$statistic, shapiro_p = sw_test$p.value)
  }
})
normality_results %>% render_table(caption = "Tests de Normalidad (Shapiro-Wilk)")
```

Table 5: Tests de Normalidad (Shapiro-Wilk)

	variable	shapiro_w	shapiro_p
W...1	age	0.9550999	0.0000000
W...2	bmi	0.9894079	0.0000000
W...3	sleep_hours	0.9989825	0.0038962
W...4	daily_steps	0.9981960	0.0000157

```
dist_suggestions <- num_extended %>% filter(variable %in% c("age","bmi","sleep_hours","daily_steps")) %
  sugerida = case_when(
    abs(skew) < 0.3 & abs(kurt) < 1 ~ "Normal",
    skew > 0.8 ~ "Gamma / Log-normal",
    skew > 0.3 & skew <= 0.8 ~ "Asimetría leve (aprox. Normal tras transf.)",
    TRUE ~ "Revisar visual"
  )
) %>% select(variable, skew, kurt, sugerida)
dist_suggestions %>% render_table(caption="Sugerencias distribucionales basadas en forma", html_font_si
```

Table 6: Sugerencias distribucionales basadas en forma

variable	skew	kurt	sugerida
age	0.0012578	-1.2012359	Revisar visual
bmi	0.4206923	0.2137246	Asimetría leve (aprox. Normal tras transf.)
daily_steps	0.0524075	-0.1437885	Normal
sleep_hours	0.0251318	-0.1227913	Normal

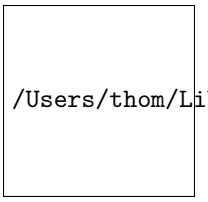
8. Análisis Mixto (Cuantitativas vs Categóricas)

```
if(all(c("bmi","gender") %in% names(df))) {
  bmi_gender_stats <- df %>% filter(!is.na(bmi), !is.na(gender)) %>% group_by(gender) %>% summarise(n =
  bmi_gender_stats %>% render_table(caption = "BMI por género")
}
```

Table 7: BMI por género

gender	n	mean	sd	median	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

```
if(all(c("bmi","gender") %in% names(df))) {
  p_box <- ggplot(df, aes(x = gender, y = bmi, fill = gender)) + geom_boxplot(alpha = 0.7) + theme_mini
  p_hist <- ggplot(df, aes(x = bmi, fill = gender)) + geom_histogram(position = "identity", alpha = 0.4
  p_box + p_hist
}
```



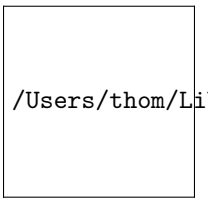

```
# Análisis adicional: Age por Target si existe
if(all(c("age","target") %in% names(df))) {
  age_target_stats <- df %>% filter(!is.na(age), !is.na(target)) %>%
    group_by(target) %>%
    summarise(n = n(), mean = mean(age), sd = sd(age), median = median(age),
              q1 = quantile(age,0.25), q3 = quantile(age,0.75), .groups = "drop")
  age_target_stats %>% render_table(caption = "Edad por Target")
}
```

Table 8: Edad por Target

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

```
# Gráficos adicionales mixtos
if(all(c("age","target") %in% names(df))) {
  p_age_box <- ggplot(df, aes(x = target, y = age, fill = target)) +
    geom_boxplot(alpha = 0.7) + theme_minimal() + theme(legend.position = "none")

  if("sleep_hours" %in% names(df)) {
    p_sleep_box <- ggplot(df, aes(x = target, y = sleep_hours, fill = target)) +
      geom_boxplot(alpha = 0.7) + theme_minimal() + theme(legend.position = "none")
    p_age_box + p_sleep_box
  } else {
    p_age_box
  }
}
```



9. Pruebas de Hipótesis

9.1 Media de BMI vs 25

```
alpha <- 0.05
if("bmi" %in% names(df)) {
  bmi_vec <- df %>% filter(!is.na(bmi)) %>% pull(bmi)
  set.seed(123)
  shapiro <- shapiro.test(sample(bmi_vec, min(5000, length(bmi_vec))))
  t_res <- t.test(bmi_vec, mu = 25)
```

```
  d_val <- effectsize::cohens_d(bmi_vec, mu = 25)
  decision <- if(t_res$p.value < alpha) "Se rechaza H0: evidencia de diferencia" else "No se rechaza H0
  render_table(
    tibble(
    Hipotesis_Nula = "mu = 25",
    Hipotesis_Alterna = "mu != 25",
    Media_Muestral = mean(bmi_vec),
    Estadistico_t = unname(t_res$statistic),
    gl = unname(t_res$parameter),
    p_valor = t_res$p.value,
    IC95_LI = t_res$conf.int[1],
    IC95_LS = t_res$conf.int[2],
    Cohens_d = d_val$Cohens_d,
    Decision = decision,
    Normalidad_Shapiro_p = shapiro$p.value
  ),
    caption = "Prueba t de una muestra para BMI",
    html_font_size = 9
  )
}
```

Table 9: Prueba t de una muestra para BMI

Hipotesis_Nula	Hipotesis_Alterna	Media_Muestral	Estadistico_t	gl	p_valor	IC95_LI	IC95_LS	Cohens_d	Decision	Normalidad_Shapiro_p
mu = 25	mu != 25	24.49388	-26.89442	99999	0	24.45692	24.53076	-0.0850476	Se rechaza H0: evidencia de diferencia	0

9.2 Proporción de “healthy” vs 0.5

```
if("target" %in% names(df)) {
  targ <- df %>% filter(!is.na(target))
  n_tot <- nrow(targ)
  n_healthy <- sum(targ$target == "healthy")
  p0 <- 0.5
  p_hat <- n_healthy/n_tot
  prop_res <- prop.test(n_healthy, n_tot, p = p0, correct = FALSE)
  cohens_h <- 2*asin(sqrt(p_hat)) - 2*asin(sqrt(p0))
  decision <- if(prop_res$p.value < 0.05) "Se rechaza H0" else "No se rechaza H0"
  render_table(
    tibble(
    Hipotesis_Nula = "p = 0.5",
    Hipotesis_Alterna = "p != 0.5",
    p_muestral = p_hat,
    Diferencia_abs = p_hat - p0,
    Estadistico_Chi2 = unname(prop_res$statistic),
    p_valor = prop_res$p.value,
    IC95_LI = prop_res$conf.int[1],
    IC95_LS = prop_res$conf.int[2],
    Cohens_h = cohens_h,
    Decision = decision
  ),
```

```
    caption="Prueba de proporción para categoría 'healthy'",
    html_font_size = 9
  )
}
```

Table 10: Prueba de proporción para categoría 'healthy'

Hipotesis_Nula	Hipotesis_Alterna	p_muestra	Diferencia_abs	Estadistico_Chi2	p_valor	IC95_LI	IC95_LS	Cohens_h	Decision
p = 0.5	p != 0.5	0.70097	0.20097	16155.58	0	0.6981247	0.7037999	0.4136345	Se rechaza H0

10. Conclusiones

Hallazgos Principales

Variables Categóricas: - Distribución equilibrada en las principales variables demográficas - Relaciones significativas entre variables categóricas (ej: género vs target)

Variables Cuantitativas:
- Variables biométricas muestran distribuciones aproximadamente normales con algunos valores atípicos - Correlaciones esperadas entre peso, altura y BMI - Sugerencias distribucionales específicas por variable

Análisis Mixto: - Diferencias en estadígrafos centrales y de dispersión por categorías - Patrones visuales claros en boxplots segmentados

Pruebas de Hipótesis: - Media de BMI vs 25: interpretar tabla (considerar magnitud de d y relevancia práctica, no solo p) - Proporción "healthy" vs 0.5: interpretar diferencia absoluta y h (magnitud pequeña/mediana/grande)

11. Sesión

```
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 v
##
## 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] treemapify_2.5.6 patchwork_1.3.2  effectsize_1.0.1 rstatix_0.7.2
```

```
##  [5] corrplot_0.95    psych_2.5.6      janitor_2.2.1    lubridate_1.9.4
##  [9] forcats_1.0.0    stringr_1.5.2    dplyr_1.1.4      purrr_1.1.0
## [13] readr_2.1.5      tidyr_1.3.1      tibble_3.3.0     ggplot2_4.0.0
## [17] tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] ggfittext_0.10.2   gtable_0.3.6       xfun_0.53          bayestestR_0.17.0
##  [5] insight_1.4.2      lattice_0.22-6     tzdb_0.5.0         vctrs_0.6.5
##  [9] tools_4.5.0        generics_0.1.4     parallel_4.5.0     datawizard_1.2.0
## [13] sandwich_3.1-1     pkgconfig_2.0.3    Matrix_1.7-3       RColorBrewer_1.1-3
## [17] S7_0.2.0           lifecycle_1.0.4    compiler_4.5.0     farver_2.1.2
## [21] mnormt_2.1.1       codetools_0.2-20   carData_3.0-5      snakecase_0.11.1
## [25] htmltools_0.5.8.1  yaml_2.3.10        Formula_1.2-5      crayon_1.5.3
## [29] pillar_1.11.0      car_3.1-3          MASS_7.3-65        abind_1.4-8
## [33] multcomp_1.4-28    nlme_3.1-168       tidyselect_1.2.1   digest_0.6.37
## [37] mvtnorm_1.3-3      stringi_1.8.7      labeling_0.4.3     splines_4.5.0
## [41] fastmap_1.2.0      grid_4.5.0         cli_3.6.5          magrittr_2.0.3
## [45] survival_3.8-3     TH.data_1.1-4      broom_1.0.10       withr_3.0.2
## [49] scales_1.4.0       backports_1.5.0    bit64_4.6.0-1      timechange_0.3.0
## [53] estimability_1.5.1 rmarkdown_2.29     emmeans_1.11.2-8   bit_4.6.0
## [57] zoo_1.8-14         hms_1.1.3          coda_0.19-4.1      evaluate_1.0.5
## [61] knitr_1.50         parameters_0.28.2  mgcv_1.9-1         rlang_1.1.6
## [65] xtable_1.8-4       glue_1.8.0         vroom_1.6.5        R6_2.6.1
```