Índices ecológicos

Santos G

Tabla de contenidos

1 Contexto del proyecto

El presente documento guía describe el dataset palmerpenguins (mediciones morfométricas y metadatos de tres especies de pingüinos: *Adelie*, *Chinstrap* y *Gentoo*). Antes de calcular índices ecológicos o métricas de biodiversidad, es fundamental evaluar la calidad y consistencia de los datos. Para ello se realizarán los siguientes pasos:

1. Auditar rápidamente la base (nombres, estructura, NA, duplicados, categorías).
2. Tomar decisiones prácticas y reproducibles sobre NA y duplicados (eliminar, imputar o marcar).
3. Marcar y documentar outliers sin eliminarlos automáticamente; preparar la tabla limpia para análisis ecológicos.
4. Generar una matriz de abundancias / comunidad mínima (p. ej. conteos por isla × especie) para calcular índices y rarefacción.

2 Limpieza y revisión inicial de los datos

```
#|label: prep
#Librerías
library(tidyverse)
library(janitor)
library(skimr)
library(palmerpenguins)

df_raw <- penguins %>% as_tibble() # guardo raw para auditoría
```

```
#|label: clean_names
df <- df_raw %>% clean_names()
names(df) # comprobar
```

```
[1] "species"           "island"            "bill_length_mm"
[4] "bill_depth_mm"     "flipper_length_mm" "body_mass_g"
[7] "sex"               "year"
```

```
#|label: skim
skim(df) #Resumen rápido (estructura + NA)
```

Tabla 1: Data summary

Name	df
Number of rows	344
Number of columns	8
_____	
Column type frequency:	
factor	3
numeric	5
_____	
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
species	0	1.00	FALSE	3	Ade: 152, Gen: 124, Chi: 68
island	0	1.00	FALSE	3	Bis: 168, Dre: 124, Tor: 52
sex	11	0.97	FALSE	2	mal: 168, fem: 165

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
bill_length_mm	2	0.99	43.92	5.46	32.1	39.23	44.45	48.5	59.6	□□□□□
bill_depth_mm	2	0.99	17.15	1.97	13.1	15.60	17.30	18.7	21.5	□□□□□
flipper_length_mm	2	0.99	200.92	14.06	172.0	190.00	197.00	213.0	231.0	□□□□□
body_mass_g	2	0.99	4201.75	801.95	2700.0	3550.00	4050.00	4750.0	6300.0	□□□□□
year	0	1.00	2008.03	0.82	2007.0	2007.00	2008.00	2009.0	2009.0	□□□□□

La base de datos `penguins` contiene 344 registros y 8 variables, de las cuales 3 son categóricas (`species`, `island`, `sex`) y 5 numéricas (`bill_length_mm`, `bill_depth_mm`, `flipper_length_mm`, `body_mass_g`, `year`).

En las variables categóricas:

- `species`: tres categorías (*Adelie* = 152, *Gentoo* = 124, *Chinstrap* = 68), sin valores faltantes.
- `island`: tres categorías (*Biscoe* = 168, *Dream* = 124, *Torgersen* = 52), sin valores faltantes.
- `sex`: dos categorías (*male* = 168, *female* = 165) con 11 valores faltantes (3%).

En las variables numéricas:

- `bill_length_mm`: media ≈ 43.9 mm, rango 32.1–59.6 mm, con 2 valores faltantes.

- `bill_depth_mm`: media ≈ 17.2 mm, rango 13.1–21.5 mm, con 2 faltantes.
- `flipper_length_mm`: media ≈ 200.9 mm, rango 172–231 mm, con 2 faltantes.
- `body_mass_g`: media ≈ 4201 g, rango 2700–6300 g, con 2 faltantes.
- `year`: muestreos entre 2007–2009, sin valores faltantes.

Interpretación general:

- Los nombres de variables ya están estandarizados.
- El dataset presenta un bajo porcentaje de NA (<1% en mediciones y 3% en `sex`). Estos casos podrán eliminarse o imputarse según el análisis.
- No se observan inconsistencias de escritura en categorías ni rangos numéricos irreales.

```
#|label: duplicates
n_dup <- sum(duplicated(df))
n_dup # Número de filas duplicadas
```

```
[1] 0
```

```
# si quieres, ver filas duplicadas (opcional)
df %>% filter(duplicated(.)) %>% head()
```

```
# A tibble: 0 x 8
# i 8 variables: species <fct>, island <fct>, bill_length_mm <dbl>,
#   bill_depth_mm <dbl>, flipper_length_mm <int>, body_mass_g <int>, sex <fct>,
#   year <int>
```

```
# eliminar duplicados exactos (opcional)
df <- df %>% distinct()
```

En este caso no se encontraron filas duplicadas (`n_dup = 0`).

Interpretación general:

- Cuando no hay duplicados, no se requieren cambios.
- Si en futuros proyectos aparecen duplicados, se recomienda, verificar primero si son errores de registro o réplicas biológicas válidas, ya que si son errores (mismo individuo registrado más de una vez), deben eliminarse, caso contrario deben mantenerse o promediarse según el objetivo del estudio.

```
#Variables categóricas
df %>% summarise(across(where(is.character), n_distinct))
```

```
# A tibble: 1 x 0
```

```
df %>% summarise(across(where(is.factor), n_distinct))
```

```
# A tibble: 1 x 3
  species island   sex
    <int>  <int> <int>
1       3      3     3
```

```
df %>% count(species)
```

```
# A tibble: 3 x 2
  species       n
  <fct>     <int>
1 Adelie      152
2 Chinstrap    68
3 Gentoo      124
```

```
df %>% count(island)
```

```
# A tibble: 3 x 2
  island        n
  <fct>     <int>
1 Biscoe      168
2 Dream       124
3 Torgersen    52
```

```
df %>% count(sex)
```

```
# A tibble: 3 x 2
  sex        n
  <fct>  <int>
1 female   165
2 male     168
3 <NA>      11
```

```
#Tipos de variables (convertir a factor si hace falta)
df <- df %>%
  mutate(species = as.factor(species),
         island  = as.factor(island),
         sex     = as.factor(sex))

#Revisar posibles inconsistencias de texto
unique(df$species)
```

```
[1] Adelie    Gentoo    Chinstrap
Levels: Adelie Chinstrap Gentoo
```

```
unique(df$island)
```

```
[1] Torgersen Biscoe    Dream
Levels: Biscoe Dream Torgersen
```

```r
unique(df$sex)
```

```
[1] male   female <NA>
Levels: female male
```

```r
df1 <- df %>% select(-year) #Eliminar columnas innecesarias
```

Las variables categóricas (`species`, `island`, `sex`) presentan 3, 3 y 2 categorías respectivamente, sin inconsistencias de escritura. Se ajustaron los tipos de variable a `factor` y se eliminó la columna `year`, ya que no será utilizada en los análisis posteriores.

```r
# Boxplots rápidos para ver rangos y "outliers" visuales
key_vars <- c("bill_length_mm","bill_depth_mm","flipper_length_mm","body_mass_g")

df %>%
  select(all_of(key_vars)) %>%
  pivot_longer(everything(), names_to = "variable", values_to = "value") %>%
  ggplot(aes(x = variable, y = value)) +
  geom_boxplot() +
  coord_flip() +
  labs(title = "Boxplots rápidos: rangos y valores extremos",
       x = "", y = "Valor") +
  theme_minimal(base_size = 13) +
  theme(
   plot.title = element_text(hjust = 0.5, face = "bold"),
   legend.position = "none",
   strip.text = element_text(face = "bold")
   )
```

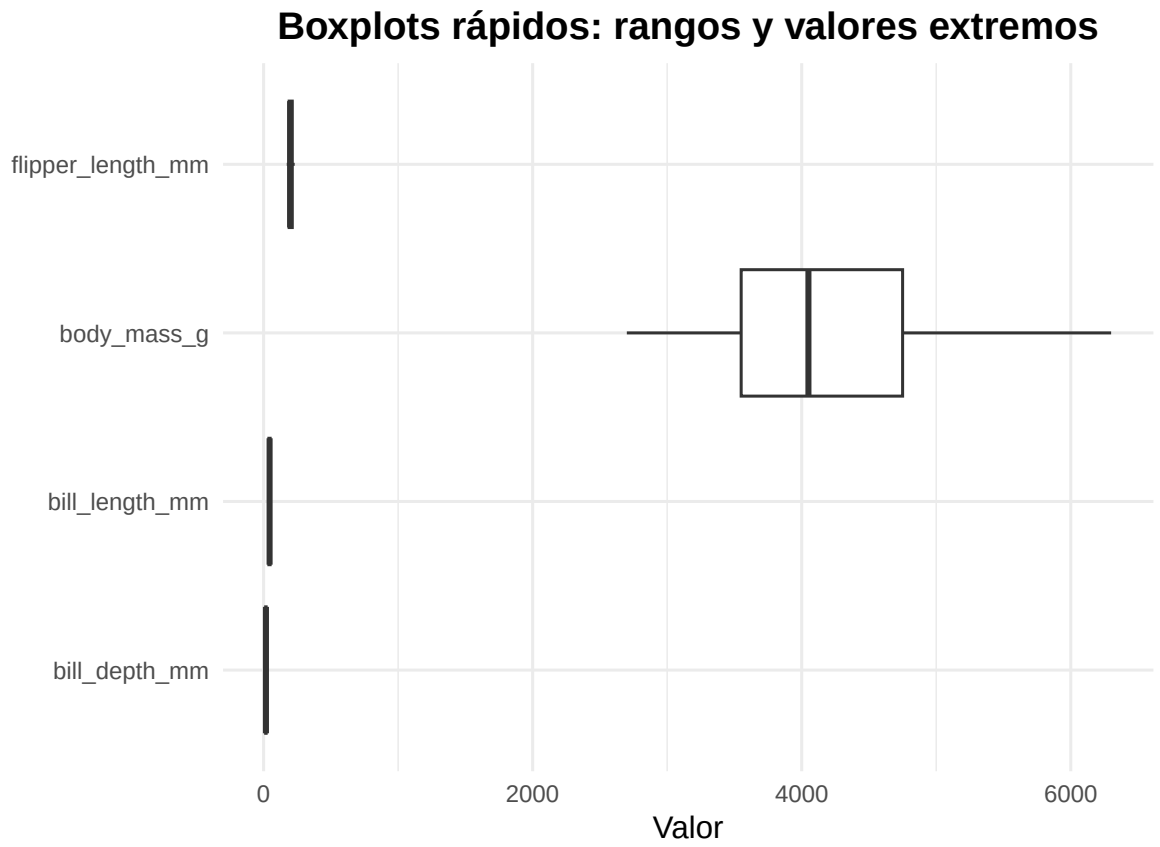

Figura 1: Rangos y valores extremos de variables morfométricas.

Los rangos numéricos se encuentran dentro de lo esperado para las especies registradas, y los gráficos de caja confirman que los valores extremos corresponden a variabilidad natural más que a errores de registro.

```
#|label: Manejo de los NA

# Manejo de los NA
df <- df %>%
  mutate(
    n_na_numeric = rowSums(is.na(select(., bill_length_mm, bill_depth_mm,
                                        flipper_length_mm, body_mass_g)))
  )

df <- df %>%
  mutate(
    sex = case_when(
      is.na(sex) & n_na_numeric <= 1 ~ "Unknown",  # casi toda la info presente
      TRUE ~ as.character(sex)                     # dejar como está
    )
  )

df <- df %>%
  filter(!(is.na(bill_length_mm) &
           is.na(bill_depth_mm) &
```

```
            is.na(flipper_length_mm) &
            is.na(body_mass_g)))


df <- df %>% select(-n_na_numeric) #Eliminar columnas innecesarias
```

Se contabilizó cuántas mediciones morfométricas tiene cada registro (bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g).

- Registros con ≥ 3 mediciones y sex = NA fueron etiquetados como sex = “Unknown”: son registros con información morfométrica suficiente como para conservarlos en análisis de biodiversidad/morfometría, pero sin identificación sexual.
- Registros con 0 mediciones (probablemente avistamientos sin mediciones) fueron eliminados, ya que no aportan datos morfométricos para los análisis previstos.