Análisis exploratorios

Santos G

Tabla de contenidos


```
# Librerías
library(tidyverse)   # Manipulación de datos: dplyr, tidyr, readr
library(janitor)     # Limpieza: clean_names(), tabyl()
library(ggplot2)     # Gráficos profesionales
library(skimr)       # EDA rápido y completo (skim())
library(GGally)      # Matriz de gráficos para variables múltiples
library(knitr)       # Tablas en Quarto
library(kableExtra)  # Tablas formateadas para informes
```

1 Contexto del proyecto

Se realizó una exploración y control de calidad de los datos de entrada para identificar variables relevantes, evaluar supuestos básicos y priorizar rutas analíticas. El objetivo es generar una guía reproducible que permita a futuros analistas (o a un equipo de consultoría) replicar y ampliar los análisis según objetivos específicos (p. ej. comparar tratamientos, modelar abundancias o construir índices de condición).

2 Carga y verificación inicial de datos

```
#|label: data-load
# Carga de datos (ejemplo iris) y limpieza mínima
data("iris")
df <- as_tibble(iris) %>%
  janitor::clean_names()   # convierte a snake_case: sepal_length, etc.

# Información básica
```

```r
n_rows <- nrow(df); n_cols <- ncol(df)
glimpse(df)
```

```
Rows: 150
Columns: 5
$ sepal_length <dbl> 5.1, 4.9, 4.7, 4.6, 5.0, 5.4, 4.6, 5.0, 4.4, 4.9, 5.4, 4.~
$ sepal_width  <dbl> 3.5, 3.0, 3.2, 3.1, 3.6, 3.9, 3.4, 3.4, 2.9, 3.1, 3.7, 3.~
$ petal_length <dbl> 1.4, 1.4, 1.3, 1.5, 1.4, 1.7, 1.4, 1.5, 1.4, 1.5, 1.5, 1.~
$ petal_width  <dbl> 0.2, 0.2, 0.2, 0.2, 0.2, 0.4, 0.3, 0.2, 0.2, 0.1, 0.2, 0.~
$ species      <fct> setosa, setosa, setosa, setosa, setosa, setosa, setosa, s~
```

```r
skim(df) # Resumen compacto por variable
```

Tabla 1: Data summary

Name	df
Number of rows	150
Number of columns	5
_____	
Column type frequency:	
factor	1
numeric	4
_____	
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
species	0	1	FALSE	3	set: 50, ver: 50, vir: 50

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
sepal_length	0	1	5.84	0.83	4.3	5.1	5.80	6.4	7.9	
sepal_width	0	1	3.06	0.44	2.0	2.8	3.00	3.3	4.4	
petal_length	0	1	3.76	1.77	1.0	1.6	4.35	5.1	6.9	
petal_width	0	1	1.20	0.76	0.1	0.3	1.30	1.8	2.5	

El dataset contiene **N = 150** observaciones y **5** variables. Las variables cuantitativas son: `sepal_length`, `sepal_width`, `petal_length`, `petal_width` (continuas, en cm). La variable categórica `species` indica tres grupos balanceados (n = 50 por grupo). No se detectaron valores faltantes ni duplicados tras una inspección inicial. Esta estructura (muestras balanceadas y variables continuas sin NA) permite aplicar análisis univariados, comparativos y multivariados con mínima preprocesamiento.

La variable **`sepal_length`** presenta una media **5.84 cm**, desviación estándar (SD) **0.83 cm**, rango 4.3–7.9.

La variable **sepal_width** presenta una media **3.06 cm**, SD **0.44 cm**, rango 2.0–4.4.

La variable **petal_length** presenta una media media **3.76 cm**, SD **1.77 cm**, rango 1.0–6.9.

La variable **petal_width** presenta una media **1.20 cm**, SD **0.76 cm**, rango 0.1–2.5.

Además, el conjunto de datos presenta las siguientes características:

- **Escala homogénea de medidas**: todas las variables cuantitativas están en la misma unidad (centímetros), lo que facilita comparaciones directas y análisis multivariados sin necesidad inmediata de reescalado.
- **Posible colinealidad**: inspecciones preliminares sugieren que `petal_length` y `petal_width` están altamente correlacionadas, lo que puede condicionar modelos de regresión o técnicas de reducción de dimensionalidad (ej. PCA).
- **Grupos biológicos claros**: las especies representan categorías naturales y balanceadas, condición poco común en datos ecológicos reales donde suele haber desbalance → este dataset es ideal como caso didáctico y de referencia.
- **Potencial de discriminación**: dado el balance de clases y la separación conocida entre especies de Iris, el dataset es adecuado para ilustrar desde comparaciones descriptivas hasta modelos predictivos (GLM, LDA, clustering).

3 Matriz de correlaciones y distribuciones entre variables numéricas

```
num_df <- df %>% select(where(is.numeric))
GGally::ggpairs(num_df, upper = list(continuous = wrap("cor", size = 3)))
```

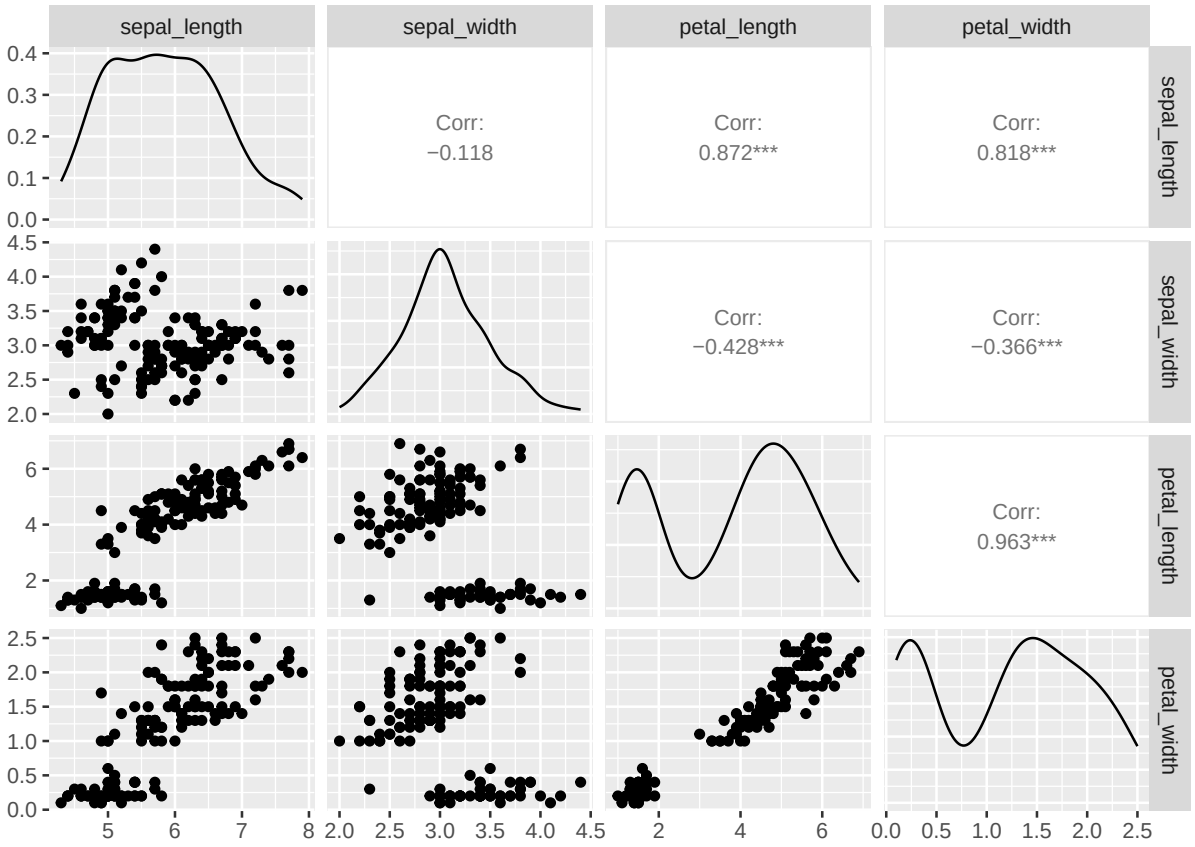

Figura 1: Matriz de dispersión y correlación de las variables cuantitativas.

La matriz de gráficos (scatterplots + histogramas + coeficientes de correlación) permite evaluar de un vistazo correlaciones, distribuciones marginales y relaciones lineales o no lineales entre variables.

- **Colinealidad fuerte**: `petal_length` y `petal_width` muestran correlación muy alta (r > 0.9), lo que evidencia redundancia informativa.
- **Correlaciones moderadas**: `sepal_length` correlaciona de forma consistente con `petal_length` (r 0.87), lo que sugiere que ambas crecen conjuntamente y pueden estar asociadas a un **gradiente de tamaño corporal general**.
- **Menor asociación**: `sepal_width` se muestra menos correlacionada con las demás, lo que indica que puede aportar información diferenciada en análisis multivariados.
- **Separación por especie**: en las distribuciones univariadas y bivariadas se observa que las especies tienden a formar agrupamientos diferenciados, especialmente `setosa`, lo que sugiere potencial de clasificación/discriminación.
- **Linealidad vs no linealidad**: las nubes de puntos muestran relaciones aproximadamente lineales, aunque con cierto solapamiento entre *versicolor* y *virginica*, lo cual podría requerir modelos más flexibles en análisis predictivos.
- **Distribución**: Para `sepal_length` es aproximadamente simétrica, con dispersión moderada. No aparecen valores extremos. `sepal_width` presenta una dispersión menor que en `sepal_length`; sin embargo, la distribución presenta cierta asimetría negativa (colas hacia valores pequeños). Finalmente tanto `petal_length`como `petal_width`muestran

una distribución marcadamente bimodal, porque las especies difieren mucho en esta variable. Es la de mayor varianza relativa.

En presencia de **alta colinealidad** ($r > 0.8$), considerar retener solo una variable del par altamente correlacionado si el objetivo es simplificación. Emplear técnicas multivariadas (PCA, análisis discriminante) para reducir dimensionalidad y evitar inestabilidad en modelos de regresión.

4 Distribución de variables morfométricas entre especies

```r
# Pasar el dataset a formato largo
iris_long <- df %>%
  pivot_longer(cols = -species,
               names_to = "Variable",
               values_to = "Valor")

# Gráfico unificado
ggplot(iris_long, aes(x = species, y = Valor, fill = species)) +
  geom_boxplot(outlier.shape = 21, alpha = 0.7) +
  facet_wrap(~ Variable, scales = "free_y") +
  labs(
    title = "Comparación de variables morfométricas en especies de Iris",
    x = "Especies",
    y = "Valor (cm)"
  ) +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    legend.position = "none",
    strip.text = element_text(face = "bold")
  )
```

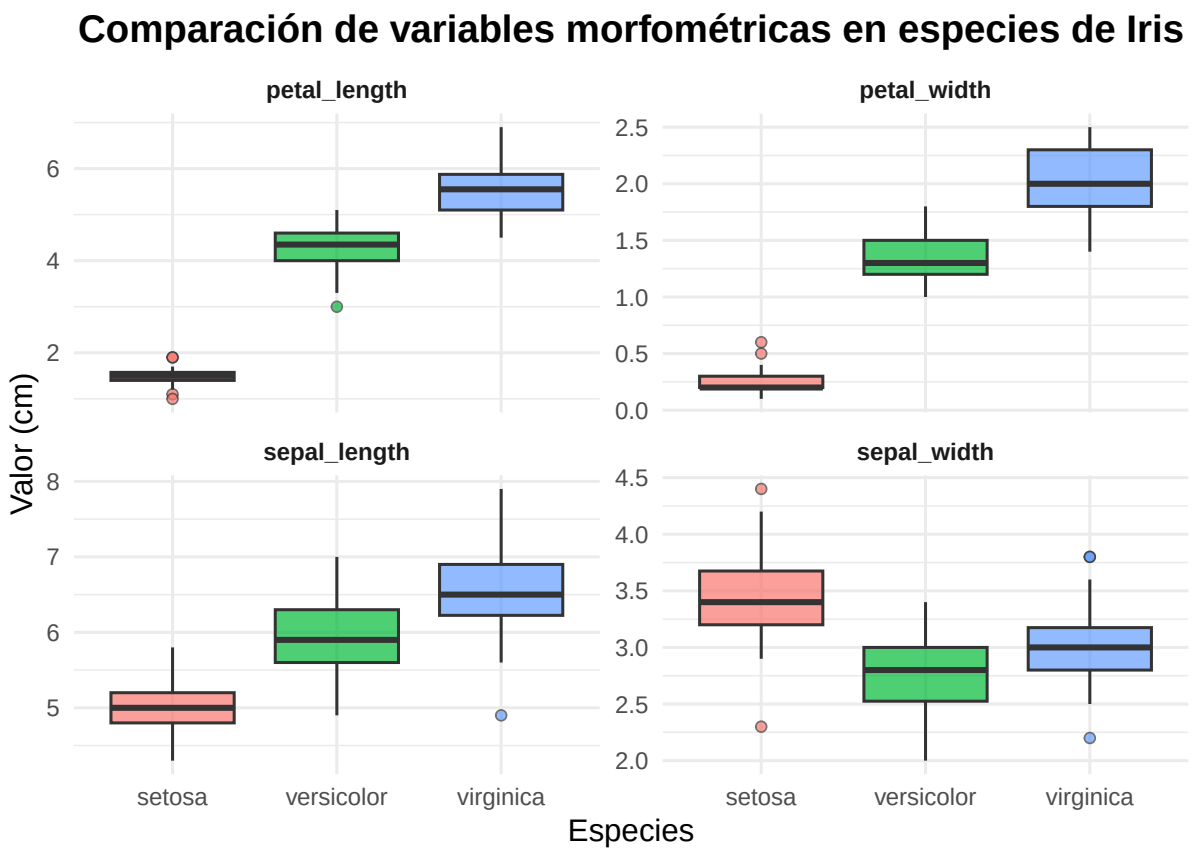

Figura 2: Distribución de variables morfométricas en tres especies de Iris.

En cuanto a la **longitud del sépalo (Sepal.Length)**, se observa que *Iris setosa* presenta los valores más bajos, con una mediana cercana a los 5 cm, claramente diferenciada de las otras especies. Por su parte, *I. versicolor* e *I. virginica* muestran longitudes mayores y con bastante solapamiento, lo que dificulta separarlas únicamente con esta variable. En consecuencia, Sepal.Length resulta útil para discriminar a *I. setosa*, pero limitado para distinguir entre *versicolor* y *virginica*.

La **anchura del sépalo (Sepal.Width)** muestra distribuciones más amplias y con solapamiento importante entre las tres especies. Aunque *setosa* tiende a tener sépalos ligeramente más anchos en promedio, la alta variabilidad reduce su capacidad discriminante. Así, esta variable por sí sola tiene bajo poder para diferenciar especies.

En el caso de la **longitud del pétalo (Petal.Length)**, se aprecia una separación casi perfecta: *setosa* forma un grupo completamente aislado con valores mucho menores, mientras que *versicolor* y *virginica* presentan medianas diferenciadas, aunque con cierto solapamiento en sus rangos intercuartílicos. Esto convierte a Petal.Length en una de las variables más informativas y clave para la clasificación.

Finalmente, la **anchura del pétalo (Petal.Width)** también resulta altamente discriminante. *Setosa* presenta valores notablemente bajos y sin superposición con las otras especies, mientras que *versicolor* y *virginica* se diferencian con claridad, aunque con menor margen que en el caso de Petal.Length. Esta variable se perfila como una de las más robustas para separar especies, especialmente en conjunto con Petal.Length.