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

El dataset contiene **N = 150 observaciones** y **5 variables**. Cuatro son cuantitativas continuas en centímetros (*Sepal.Length, Sepal.Width, Petal.Length, Petal.Width*), y una categórica (*Species*), que clasifica en tres grupos balanceados (n = 50 por especie). No se detectaron valores faltantes ni duplicados tras la inspección inicial. Esta estructura balanceada y sin NA permite aplicar análisis univariados, comparativos y multivariados con mínimo preprocesamiento.

La **Tabla 1** de estadísticos descriptivos muestra lo siguiente:

- **Sepal.Length:** media ≈ 5.84 cm, SD ≈ 0.83, rango 4.3–7.9. Variación moderada, con solapamiento esperado entre especies.
- **Sepal.Width:** media ≈ 3.06 cm, SD ≈ 0.44, rango 2.0–4.4. Es la variable más estable, aunque con ligera asimetría negativa.
- **Petal.Length:** media ≈ 3.76 cm, SD ≈ 1.77, rango 1.0–6.9. Mayor dispersión relativa, con clara separación de *setosa*.
- **Petal.Width:** media ≈ 1.20 cm, SD ≈ 0.76, rango 0.1–2.5. Alta variabilidad, con potencial de discriminación entre las tres especies.

```
#|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
n_rows <- nrow(df); n_cols <- ncol(df)
tbl1<-skim(df)
tbl1 # 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	□□□□□

Aspectos destacados del dataset:

- **Escala homogénea de medidas:** todas las variables en centímetros → comparaciones y análisis multivariados sin necesidad de reescalado inmediato.
- **Colinealidad esperada:** Petal.Length y Petal.Width muestran alta correlación, lo que debe considerarse en regresiones o PCA.
- **Grupos biológicos claros y balanceados:** un escenario ideal para aprendizaje, aunque poco frecuente en estudios ecológicos reales.
- **Potencial de discriminación:** las variables de pétalos concentran el mayor poder de separación, coherente con su relevancia funcional en la biología reproductiva de las plantas.

3 Matriz de correlaciones y distribuciones entre variables numéricas

La **Figura 1** combina tres tipos de información: distribuciones univariadas, relaciones bivariadas y correlacciones númericas.

```
num_df <- df %>% select(where(is.numeric))
Fig1<- GGally::ggpairs(
  df,
  columns = 1:4, # solo variables numéricas
  mapping = aes(color = species), # color por especie
  upper = list(continuous = wrap("cor", size = 3)),
  diag = list(continuous = wrap("densityDiag", alpha = 0.6))
)
Fig1
```

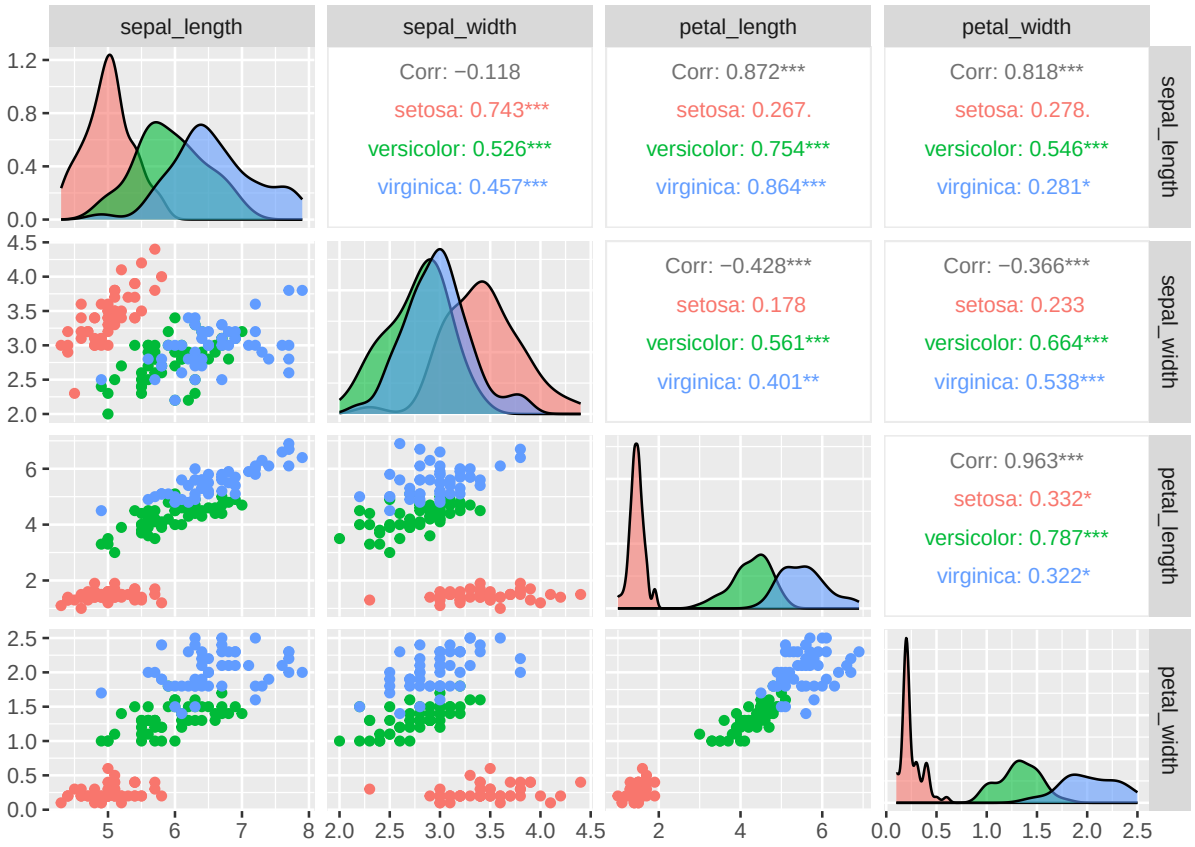

Figura 1: Matriz de dispersión y correlación de las variables cuantitativas.

3.1 Distribuciones univariadas

- **Sepal.Length**:
  - *Setosa*: concentrada en valores bajos (4.3 - 5.8 cm), muy homogénea.
  - *Versicolor*: rango intermedio (≈ 4.9 - 7.0 cm).
  - *Virginica*: valores altos (≈ 4.9 - 7.9 cm), con ligera superposición con Versicolor.
  - **Interpretación**: útil para separar *Setosa*, pero *Versicolor* y *Virginica* se solapan.
- **Sepal.Width**:
  - Distribución amplia en todas las especies.
  - *Setosa* tiende a mayores valores promedio, pero con solapamiento considerable.
  - **Interpretación**: poco poder discriminante, refleja variabilidad natural en la anchura del sépalo.
- **Petal.Length**:
  - *Setosa*: valores muy bajos (≈ 1.0 - 1.9 cm), sin solapamiento con las otras especies.
  - *Versicolor*: rango medio (≈ 3.0 - 5.1 cm).

  - *Virginica*: valores altos (≈ 4.5 - 6.9 cm).

  - **Interpretación**: variable clave, separa *Setosa* y discrimina bastante bien *Versicolor* vs *Virginica*.

- **Petal.Width**:

  - *Setosa*: valores muy bajos (≈ 0.1 - 0.6 cm).

  - *Versicolor*: rango medio (≈ 1.0 - 1.8 cm).

  - *Virginica*: valores altos (≈ 1.4 - 2.5 cm).

  - **Interpretación**: la más robusta para separar las tres especies, casi sin solapamiento.

3.2 Relaciones bivariadas

- **Sepal.Length vs Sepal.Width**: gran solapamiento entre especies, con nubes de puntos mezcladas. *Setosa* muestra ligera tendencia a sépalos más anchos.
  - **Interpretación**: baja capacidad de discriminación.
- **Sepal.Length vs Petal.Length**: patrón positivo moderado. *Setosa* queda claramente apartada (pétalos muy cortos). *Versicolor* y *Virginica* siguen una línea ascendente, con solapamiento parcial.
  - **Interpretación**: ayuda a diferenciar *Setosa*, pero no tanto entre las otras dos.
- **Sepal.Length vs Petal.Width**: tendencia positiva clara. *Setosa* aislada (pétalos estrechos). *Virginica* tiende a valores más altos.
  - **Interpretación**: más útil que *Sepal.Length* solo, pero aún con solapamientos.
- **Sepal.Width vs Petal.Length**: relación débil, nubes muy mezcladas. *Setosa* separada por bajos valores de pétalo, no por el sépalo.
  - **Interpretación**: variable *Sepal.Width* poco informativa.
- **Sepal.Width vs Petal.Width**: relación débil, con gran dispersión. *Setosa* se distingue porque tiene pétalos angostos, no por anchura del sépalo.
  - **Interpretación**: no aporta discriminación extra.
- **Petal.Length vs Petal.Width**: relación lineal muy fuerte, tres grupos claramente separados. *Setosa* aislada en valores bajos; *Versicolor* intermedia; *Virginica* en el rango alto.
  - **Interpretación**: la combinación de estas dos variables es la mejor para clasificar especies.

3.3 Correlaciones numéricas

- **Petal.Length vs Petal.Width**: r ≈ 0.96 (correlación muy alta). Variables casi redundantes, pero en conjunto definen un espacio morfológico clave.

- **Sepal.Length vs Petal.Length**: r ≈ 0.87 (correlación fuerte). A mayor sépalo, mayor pétalo, patrón general de tamaño.

- **Sepal.Length vs Petal.Width**: r ≈ 0.82 (correlación alta). También refleja el gradiente de tamaño floral.

- **Sepal.Width con el resto**: correlaciones bajas (r entre -0.4 y 0.3). Confirma que aporta poca información discriminante.

3.4 Interpretación ecológica general

- Los **pétalos** son rasgos reproductivos clave: su longitud y anchura diferencian a las especies porque están ligados a estrategias de atracción de polinizadores.

- Los **sépalos**, en cambio, son más plásticos y menos específicos, lo que explica su bajo poder discriminante.

- La **correlación entre variables de pétalo** refleja que ambas describen el mismo fenómeno biológico (tamaño floral), pero su combinación refuerza la clasificación.

- En un contexto real de ecología vegetal, esto sugiere que la diferenciación entre especies del género *Iris* depende más de rasgos reproductivos (pétalos) que de rasgos de soporte (sépalos).

4 Distribución de variables morfométricas entre especies

La **Figura 2** presenta diagramas de caja y bigotes que resumen la variación de las cuatro variables morfométricas en las tres especies de *Iris*. Este análisis complementa al resumen numérico y a la matriz de relaciones bivariadas, ya que enfatiza tendencias centrales, dispersión y presencia de datos atípicos.

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

# Gráfico unificado
Fig2 <- 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")
  )
Fig2
```

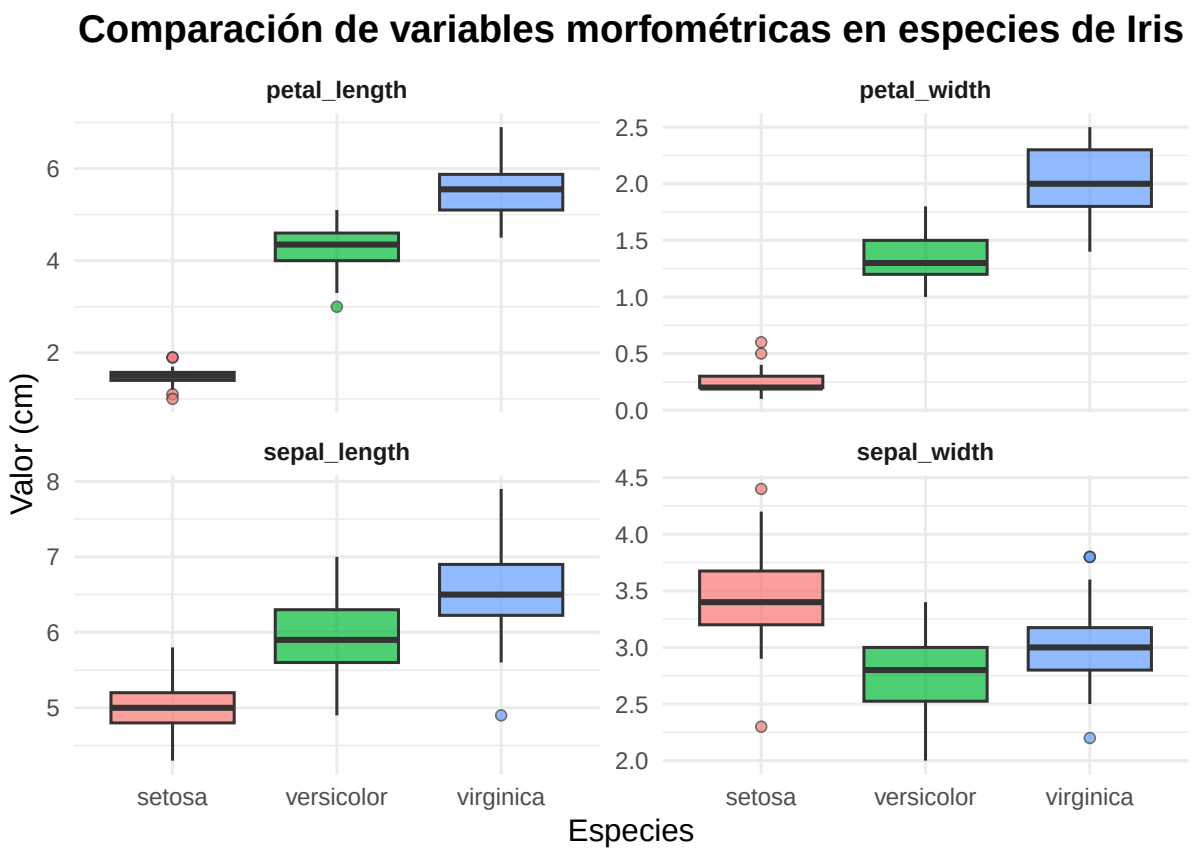

Figura 2: Distribución de variables morfométricas en tres especies de Iris.

4.1 Sepal.Length (longitud del sépalo)

- *I. setosa* muestra valores concentrados entre **4.3 y 5.8 cm**, con una mediana cercana a **5.0 cm**. La caja es compacta, indicando baja variabilidad.
- *I. versicolor* se distribuye entre **4.9 y 7.0 cm**, con mediana ≈ **5.9 cm**, rango intermedio.
- *I. virginica* alcanza los valores más altos (**4.9–7.9 cm**), con mediana ≈ **6.5 cm**.

Interpretación: hay cierto solapamiento entre *versicolor* y *virginica*. Útil para distinguir a *setosa*, pero no para discriminar con precisión entre *versicolor* y *virginica*.

4.2 Sepal.Width (anchura del sépalo)

- *I. setosa* tiene la mediana más alta (≈ **3.4 cm**), con valores entre **2.3 y 4.4 cm**.
- *I. versicolor* oscila entre **2.0 y 3.4 cm**, mediana ≈ **2.8 cm**.
- *I. virginica* se ubica entre **2.2 y 3.8 cm**, mediana ≈ **3.0 cm**.
- Se observan varios outliers tanto en *setosa* como *virginica,* individuos con sépalos inusualmente estrechos. Estos valores atípicos podrían reflejar variación intraespecífica natural o condiciones

ambientales particulares. La fuerte dispersión y el solapamiento reducen el valor discriminante de esta variable.

Interpretación: aunque setosa tiende a mayor anchura, la amplia dispersión y solapamiento hacen que esta variable tenga bajo poder discriminante.

4.3 Petal.Length (longitud del pétalo)

- *I. setosa* presenta valores muy bajos (**1.0–1.9 cm**) con mediana ≈ **1.5 cm**.
- *I. versicolor* ocupa un rango intermedio (**3.0–5.1 cm**), con mediana ≈ **4.3 cm**.
- *I. virginica* concentra los valores más altos (**4.5–6.9 cm**), mediana ≈ **5.5 cm**.

Interpretación: el solapamiento entre *versicolor* y *virginica* es reducido y se da en los límites superiores/inferiores de sus cajas. Esta variable es altamente informativa; separa completamente a *setosa* y discrimina en gran medida a *versicolor* y *virginica*.

4.4 Petal.Width (anchura del pétalo)

- *I. setosa* tiene los valores más bajos (**0.1–0.6 cm**), con mediana ≈ **0.2 cm**, sin solapamiento con las otras especies.
- *I. versicolor* se concentra entre **1.0 y 1.8 cm**, mediana ≈ **1.3 cm**.
- *I. virginica* presenta los valores más altos (**1.4–2.5 cm**), mediana ≈ **2.0 cm**.

Interpretación: junto con *Petal.Length*, constituye la variable más robusta para separar especies; su poder discriminante es muy alto y con solapamiento mínimo.

4.5 Interpretación general

Las variables de pétalo muestran diferencias netas entre especies, con cajas bien separadas. Por lo que se considera un rasgo clave para clasificación, debido a su alta capacidad de separar grupos. A diferencias de las variables de sépalo que presentan mayor dispersión y solapamiento, lo que limita su valor clasificatorio, debido a su menor capacidad diagnóstica, más influenciados por plasticidad ambiental.

El patrón confirma que los caracteres reproductivos (pétalos) son más determinantes para la diferenciación interespecífica, mientras que los vegetativos (sépalos) presentan mayor variabilidad y menor poder diagnóstico.