Análisis exploratorios

Santos G

Tabla de contenidos


```r
# 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

```r
#|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.
```

```r

# Información básica
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
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	

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.

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
```

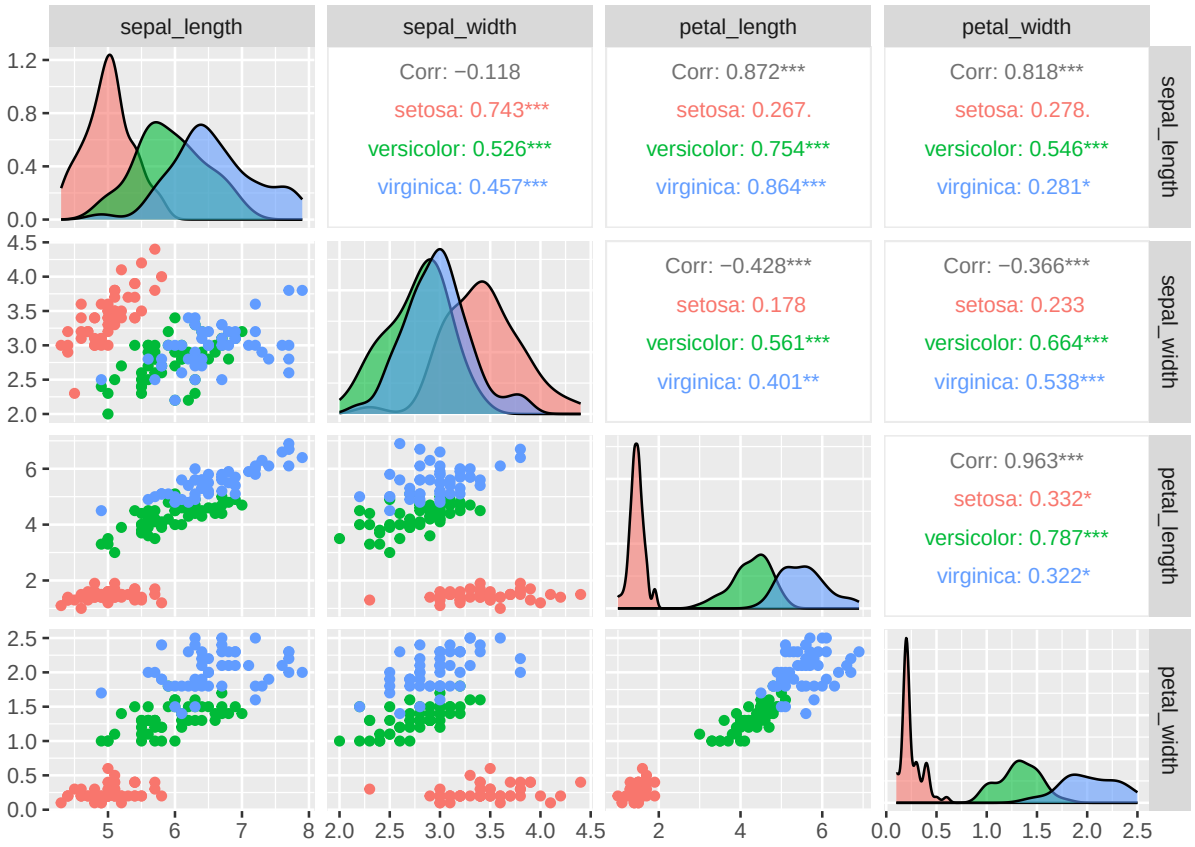

Figura 1: Matriz de dispersión y correlación de las variables cuantitativas.

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).