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(kableExtra)  # Tablas formateadas para informes
library(skimr)       # EDA rápido y completo (skim())
library(GGally)      # Matriz de gráficos para variables múltiples
library(corrplot)    # Visualización de matrices de correlación
library(knitr)       # Tablas en Quarto
```

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
# 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)
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)
```

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

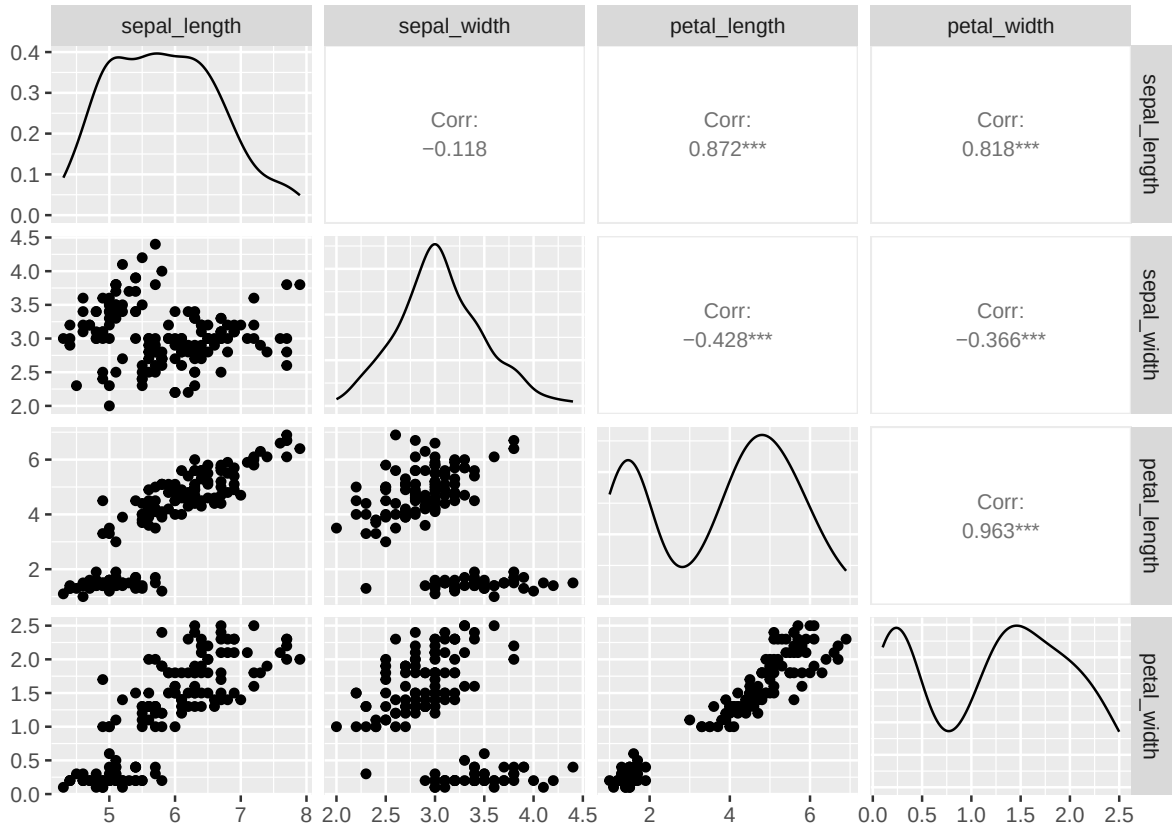
Variable type: numeric

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 $\mathbf{N = 150}$ 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.

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)))
```




```
#| label: skim-summary # Resumen compacto por variable para todo el dataset skim(df)
```