Clase 03 - Análisis exploratorio de datos

Curso Análisis de datos con R para Biociencias

Dr. José A. Gallardo | jose.gallardo@pucv.cl | Pontificia Universidad Católica de Valparaíso

19 January 2022

PLAN DE LA CLASE

1.- Introducción

- ¿Qué es un análisis exploratorio de datos?.
- ¿Por qué es importante?.
- Recomendaciones para comunicar datos de forma efectiva.

2.- Práctica con R y Rstudio cloud

- Actividad de aprendizaje ggplot2.
- Realizar un análisis exploratorio de datos.
- Realizar gráficas avanzadas con ggplot2.

ANÁLISIS EXPLORATORIO DE DATOS (EDA)

¿Qué es un análisis exploratorio de datos?
Procedimiento que permite visualizar y explorar los datos de un estudio.
¿Para qué?
1- Investigar calidad de los datos brutos.

2- Limpiar datos.

3- Observar variación de los datos.

4- Establecer un modelo básico de relación e interacción entre variables.

5- Seleccionar una prueba estadÍstica adecuada.

EDA ES UN PROCESO ITERATIVO

¿Cómo realizar un buen EDA?

1- Genera preguntas iniciales para explorar tus datos.

2- Resume, visualiza, transforma y modela tus datos.

3- Usa lo que aprendiste para generar nuevas preguntas.

Preguntas clave, pero no las únicas

- ¿Qué tipo de variación existe en la/s variables de estudio?
- ¿Qué tipo de covariación o interacción existe entre las variables de estudio?
- ¿Cuál es el modelo más simple que explica la relación entre variables?
- ¿Existen errores, datos faltantes, valores atípicos?

EDA: IMPORTANCIA DE LA ESTRUCTURA DE LOS DATOS

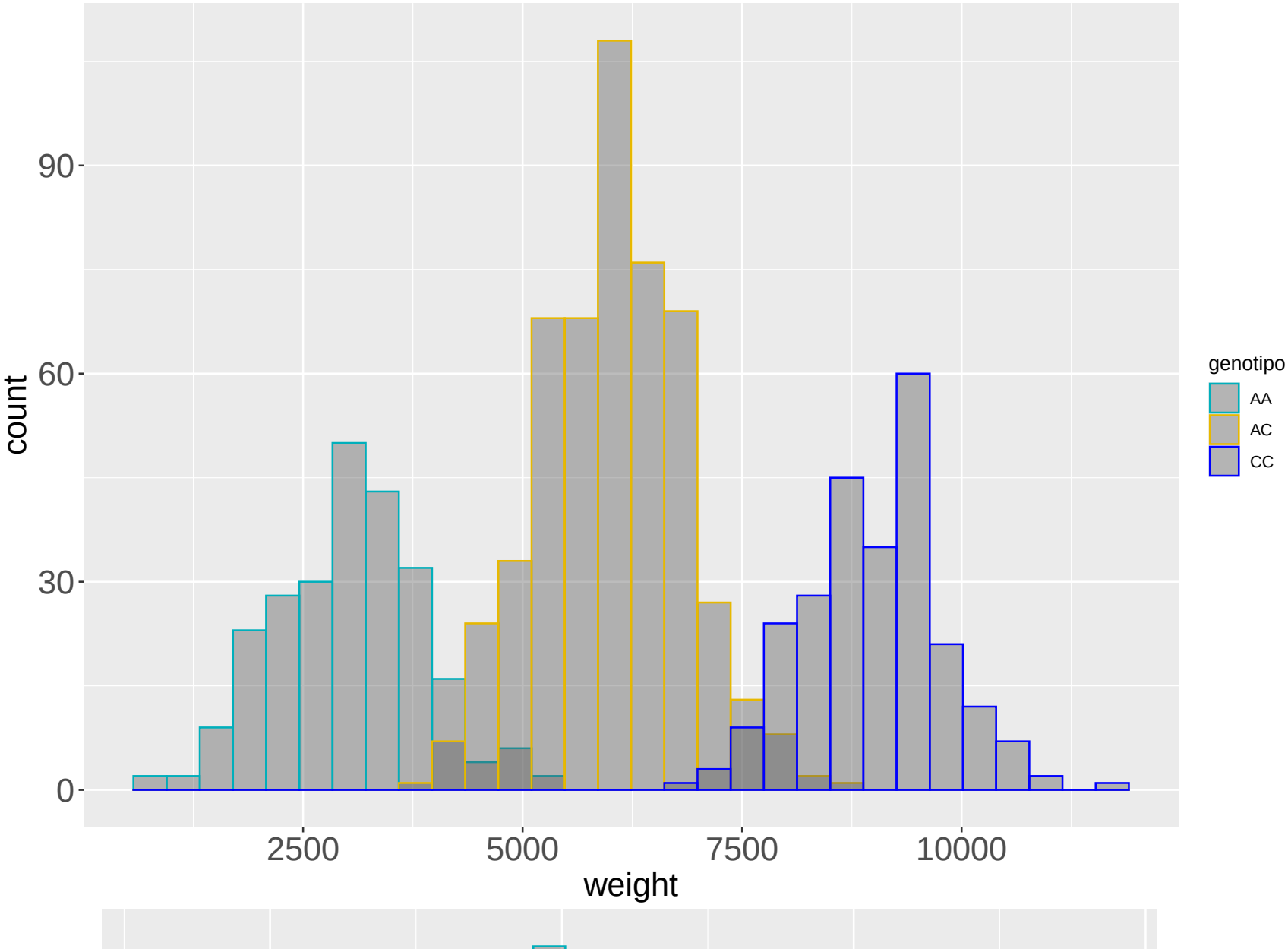
¿Mis datos son balanceados o no balanceados?

Table 1: Diseño no balanceado con datos faltantes

	d1	d2	d3	d4	d5	d6
Macho	3	3	4	2	3	0
Hembra	9	7	8	9	11	12

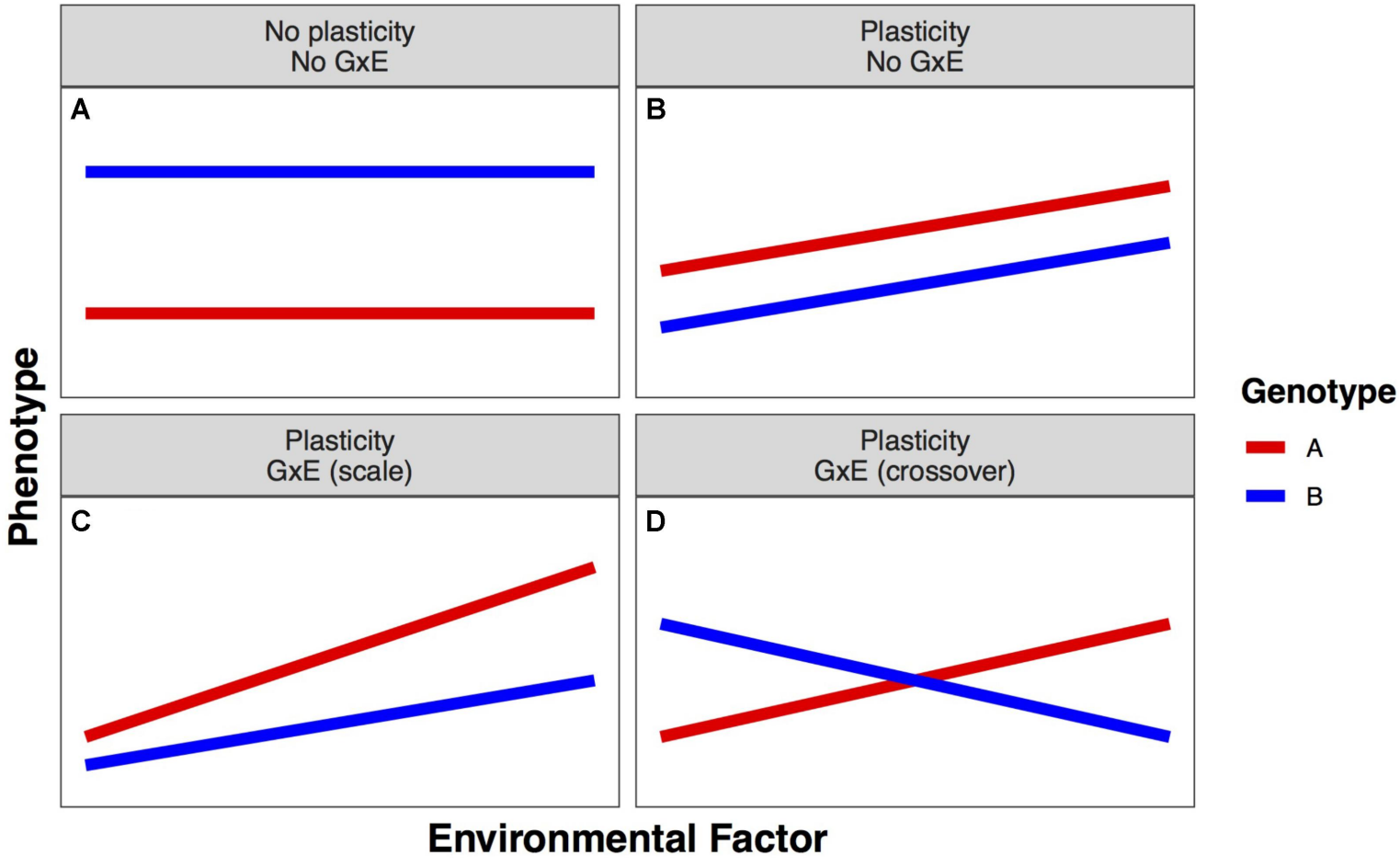
EDA: VARIACIÓN DENTRO DE UN FACTOR

¿La variación de mis datos es homogenea?



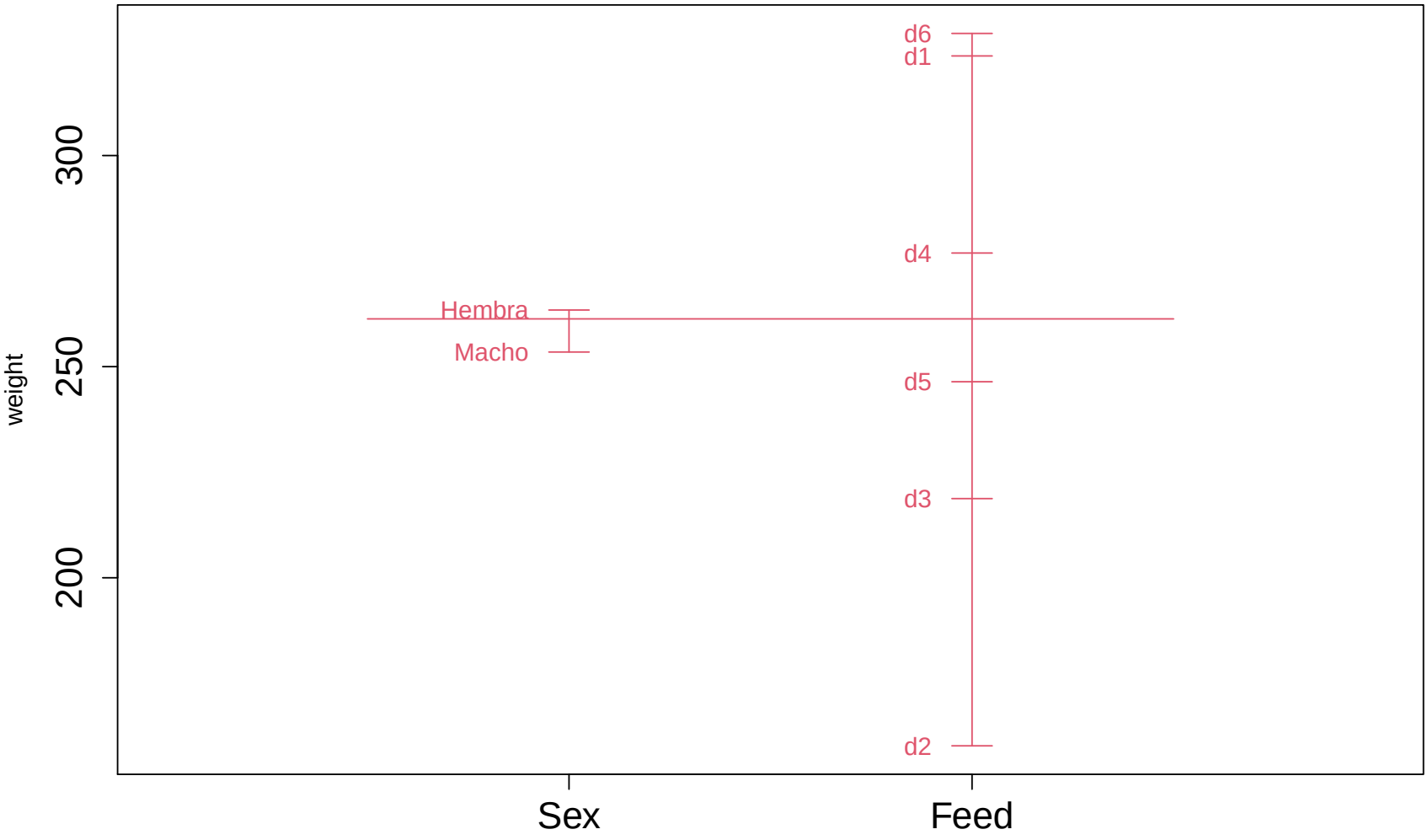
EDA: INTERACCIÓN ENTRE FACTORES

¿Existe interacción entre los factores?



EDA: TAMAÑO DE LOS EFECTOS

¿Qué factor tiene un mayor efecto sobre la variable respuesta?



EDA: CORRELACIÓN ENTRE VARIABLES CONTINUAS

¿Existe covariación / correlación entre mis datos?

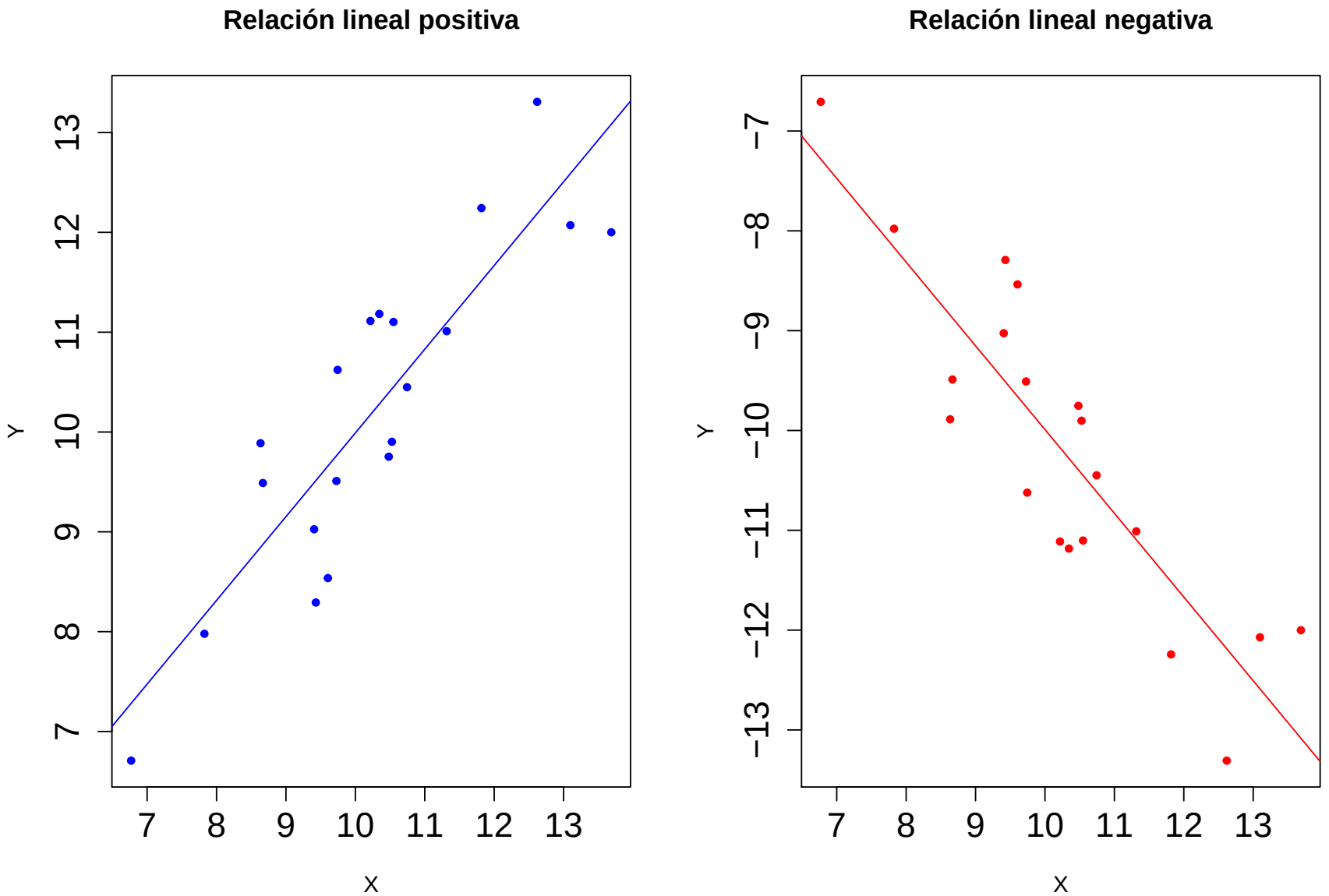
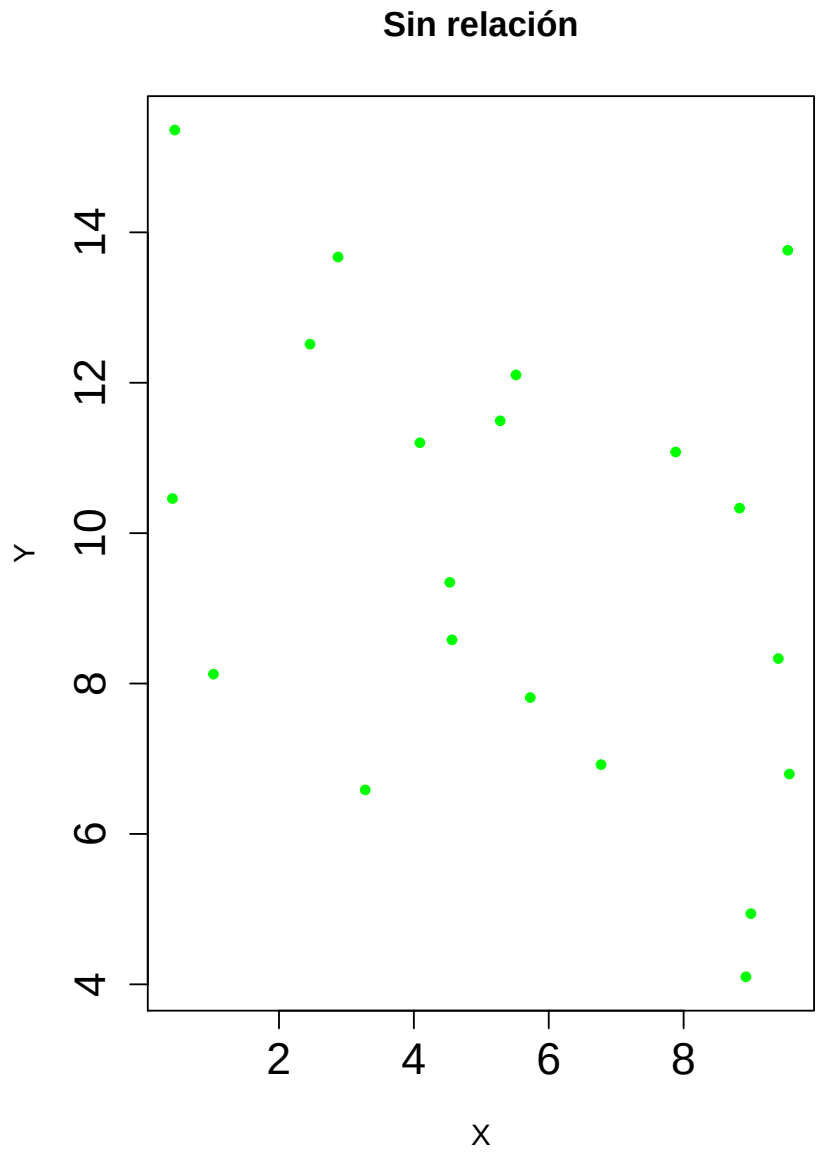
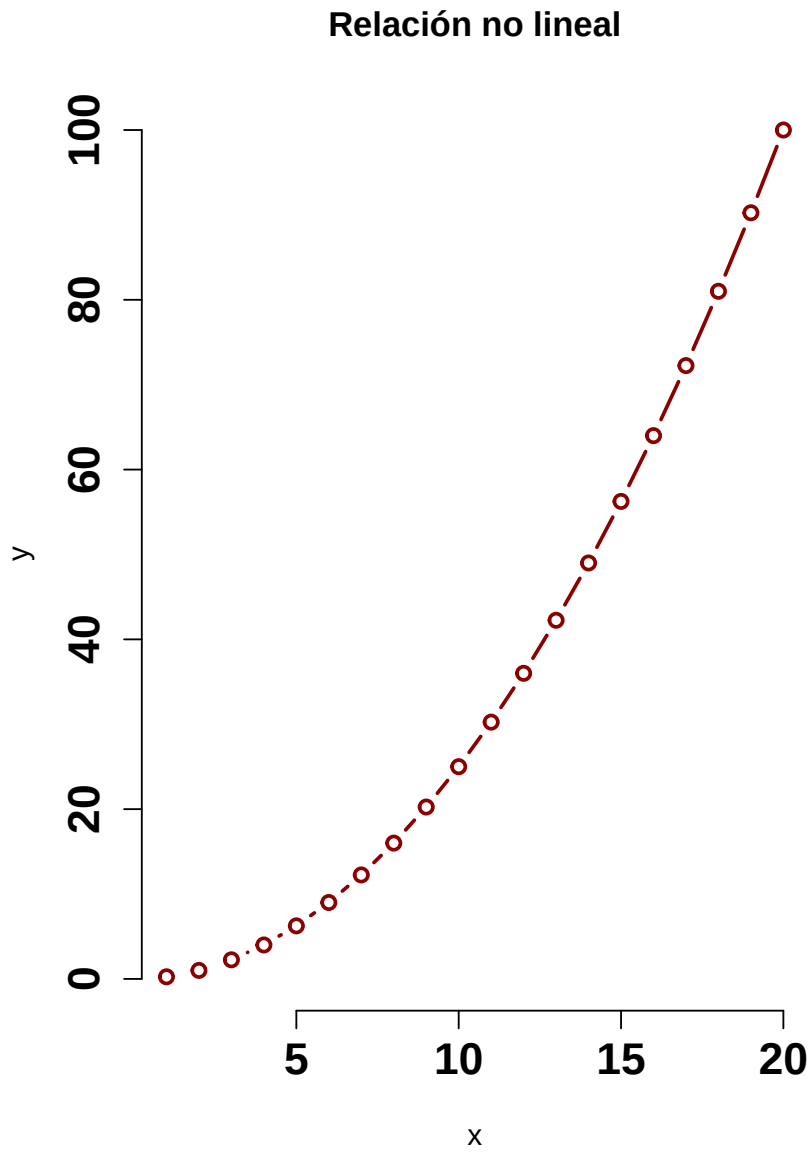


Figure 1: Distribución normal multivariada, out.width = '80%'

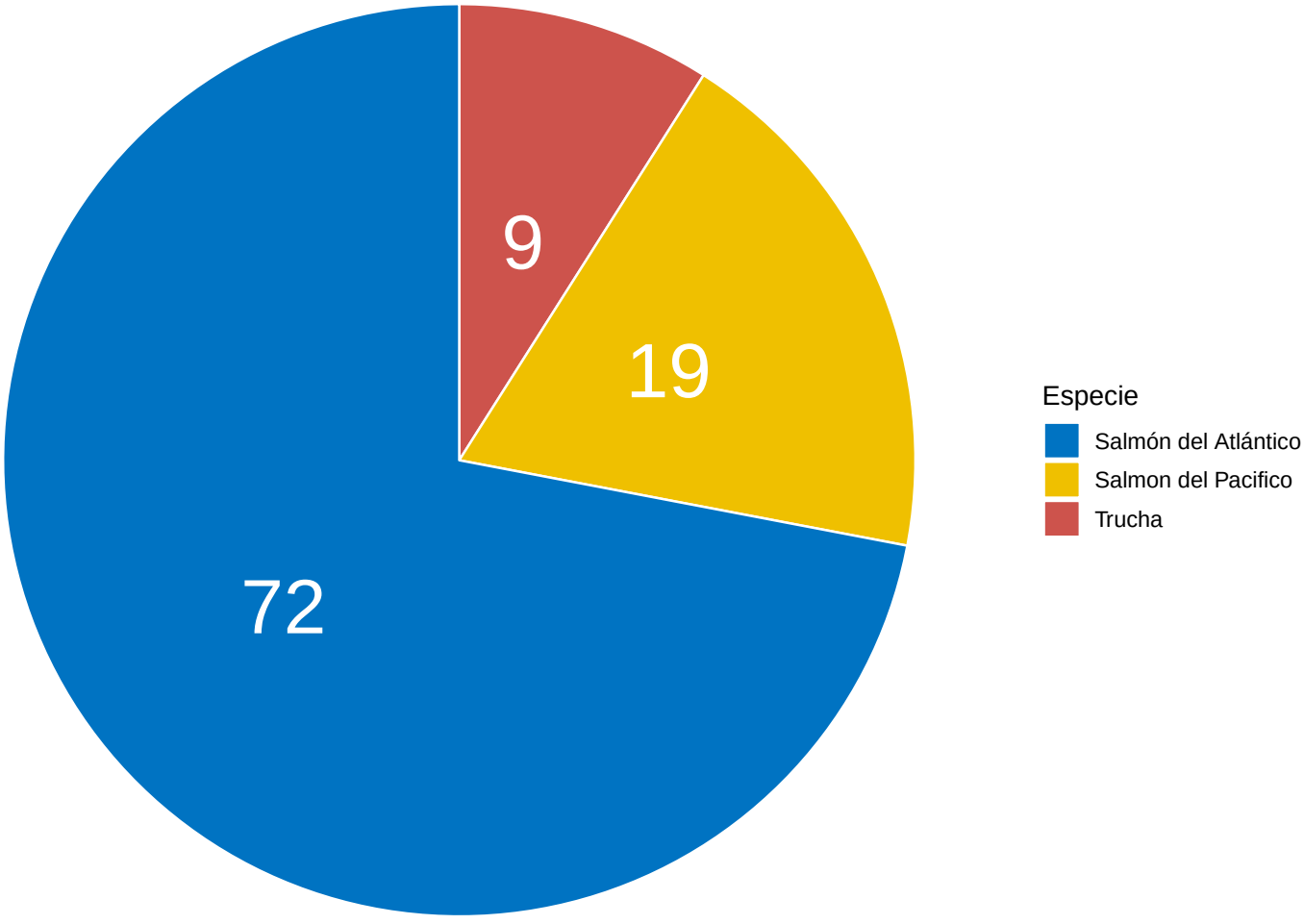
EDA: OTRAS CORRELACIONES





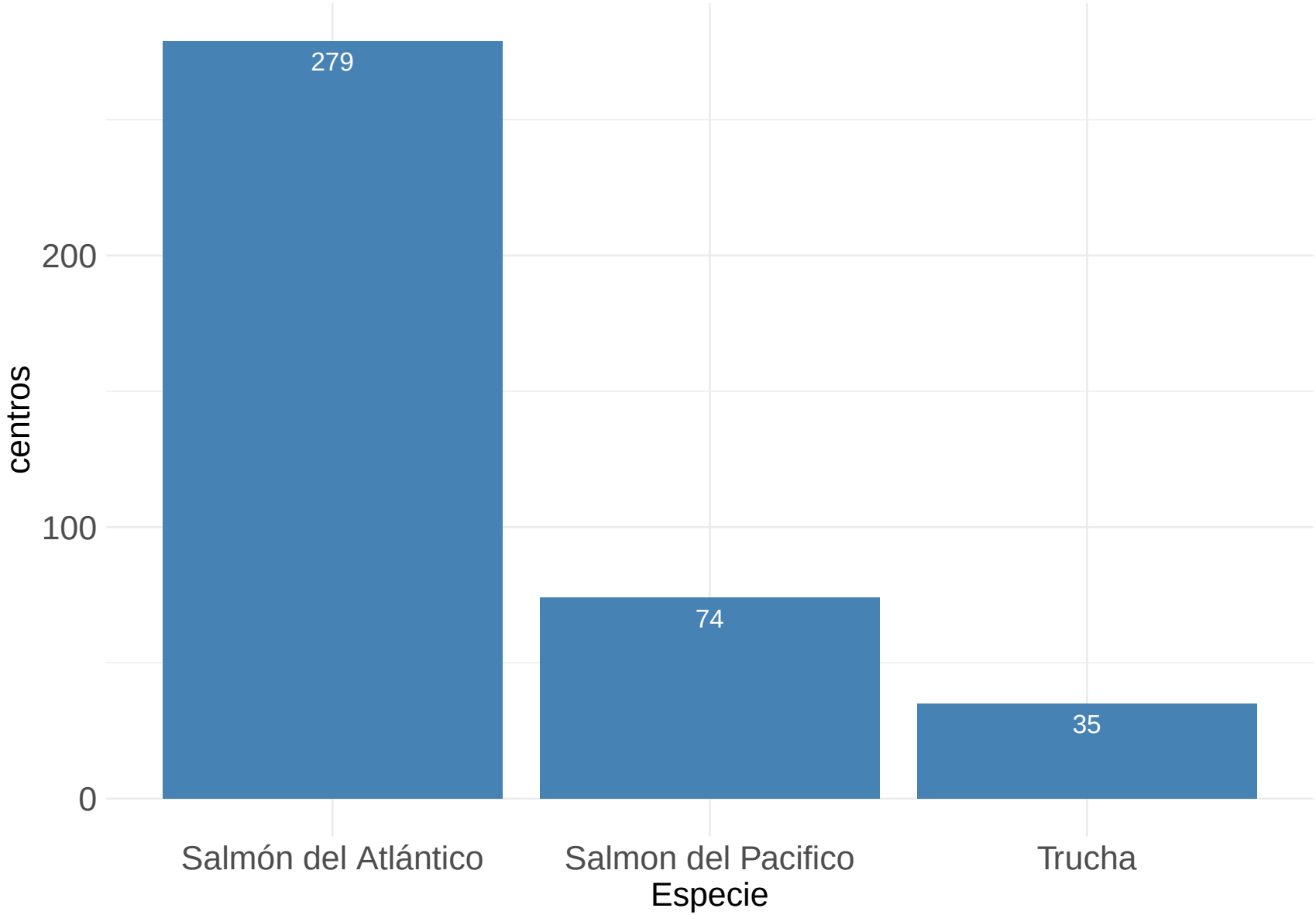
COMUNICAR EDA DE FORMA EFECTIVA

Evita las tortas



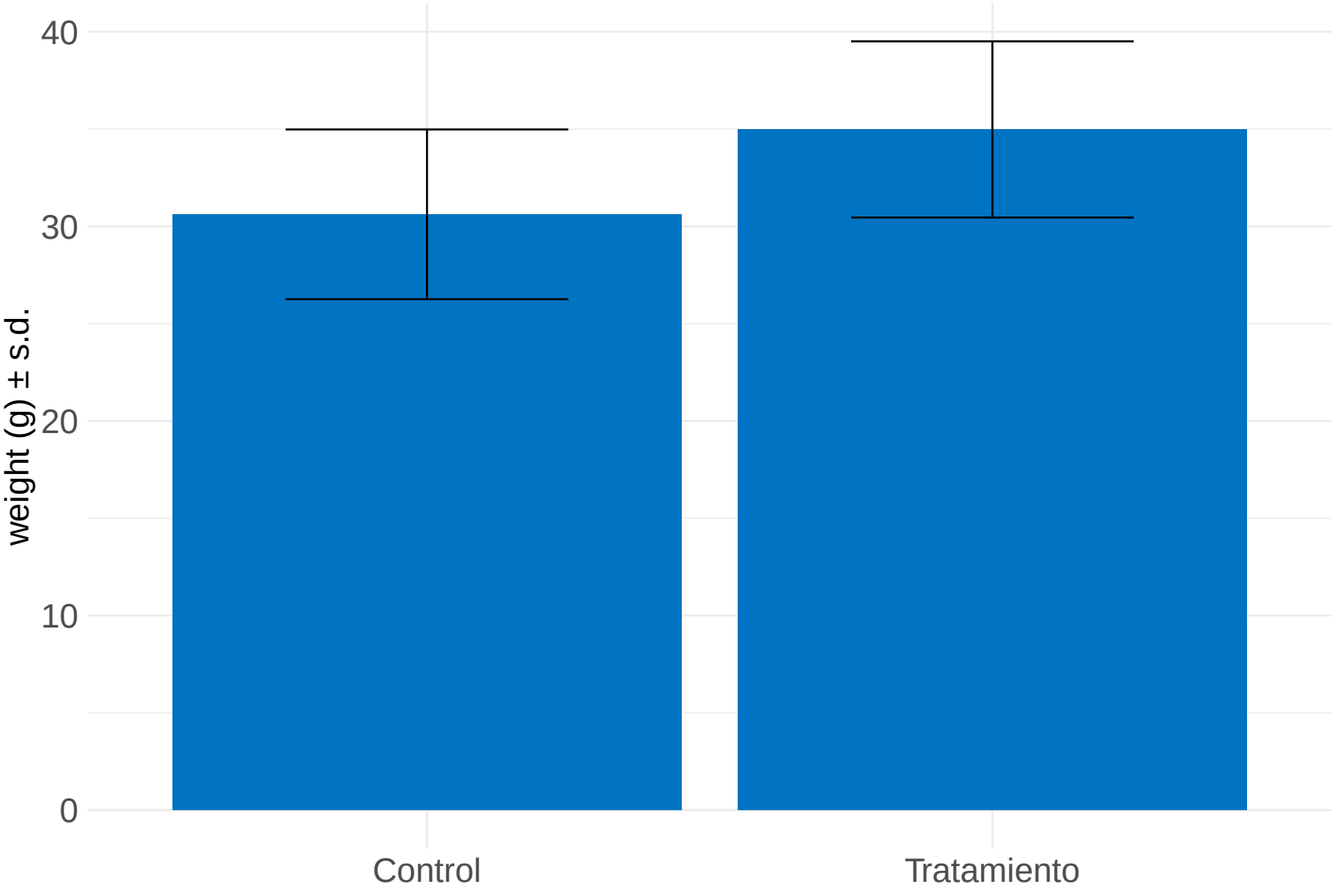
SOLUCIÓN

Prefiere datos brutos y barras



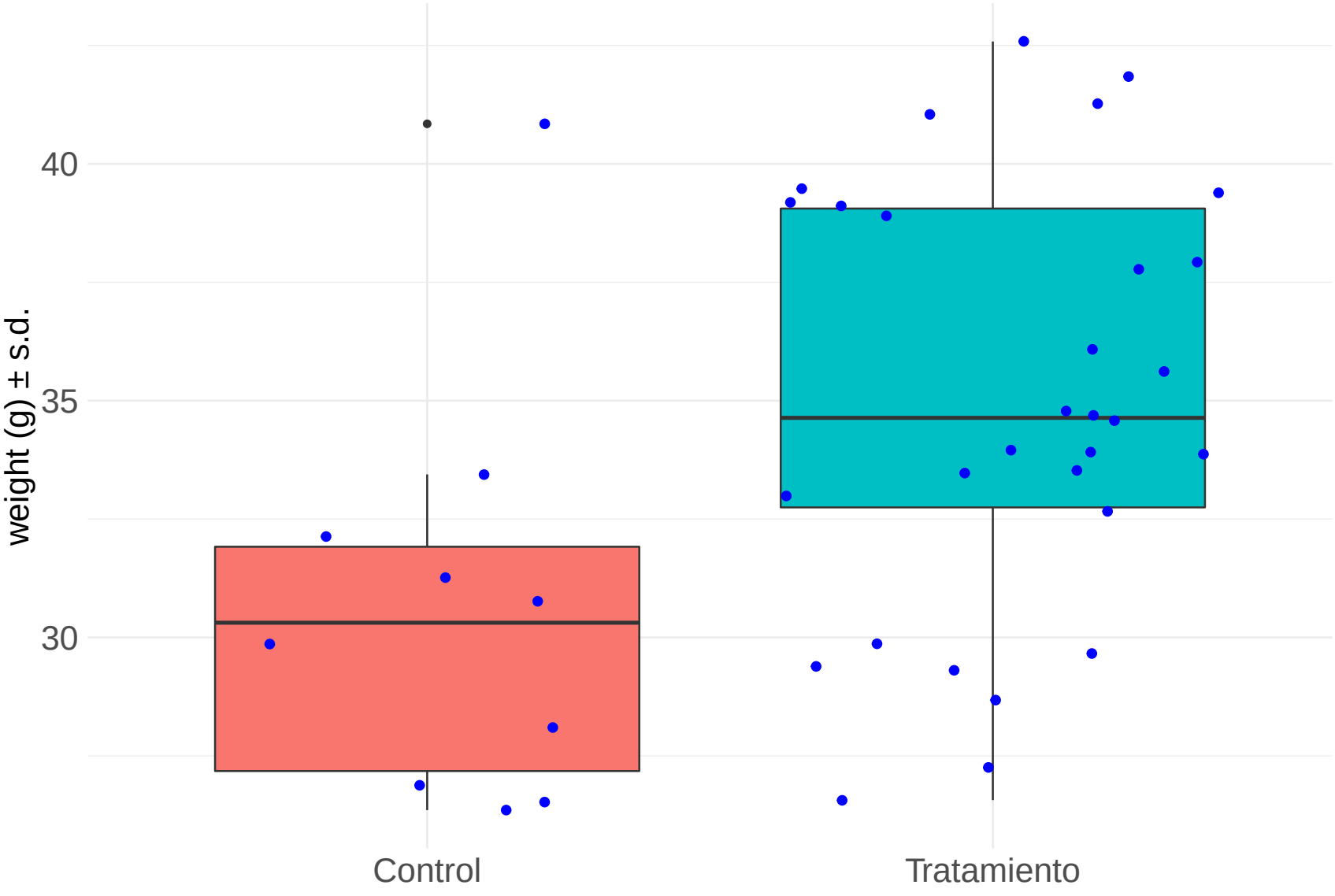
COMUNICAR EDA DE FORMA EFECTIVA 2

Evita gráficas de barras para comparar grupos



SOLUCIÓN

Prefiere boxplot, muestra tus datos ¡¡



VISUALIZACIÓN DE DATOS AVANZADO CON GGPLOT2

ggplot2: Librería de visualización de datos preferido para realizar graficas con R Wickham en 2005).
Ventajas Gran flexibilidad. Sistema para realizar gráficos completo y maduro. Una gran comunidad de desarrolladores.
Características Los datos siempre deben ser un data.frame. Usa un sistema diferente para añadir elementos al gráfico.

COMPARACIÓN GGPLOT2 - GRAPHICS

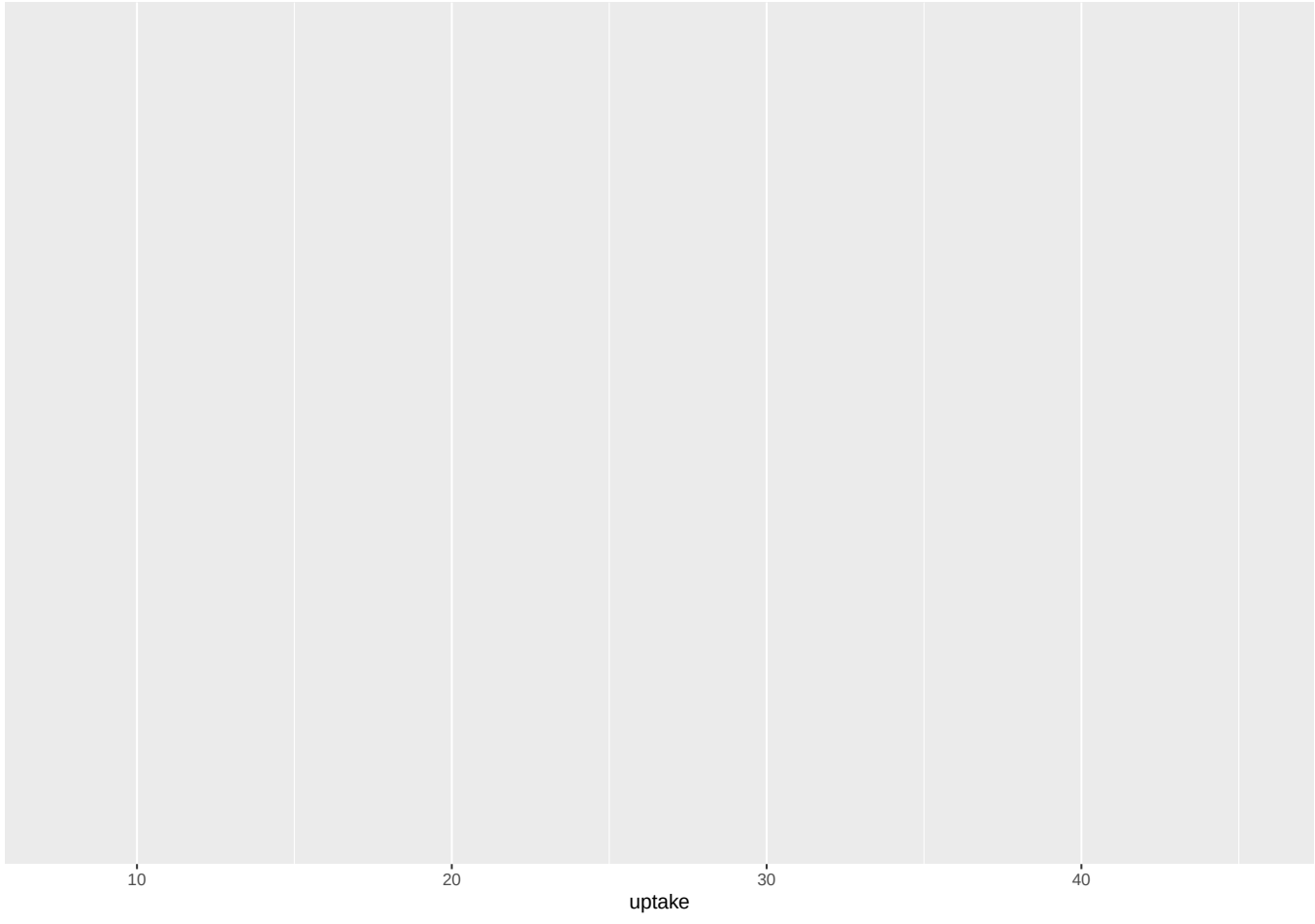
Comparación de algunos comandos de gráficas entre la librería **graphics** y **ggplot2**

Función	graphics	ggplot2
Función genérica para graficar	plot()	ggplot()
Histogramas	hist()	geom_histogram()
Gráfica de cajas y bigotes	boxplot()	geom_boxplot()
Etiquetar ejes	xlab=“ ” , ylab=“ ”	labs(x=““, y=””“)

¿CÓMO FUNCIONA GGPLOT2?

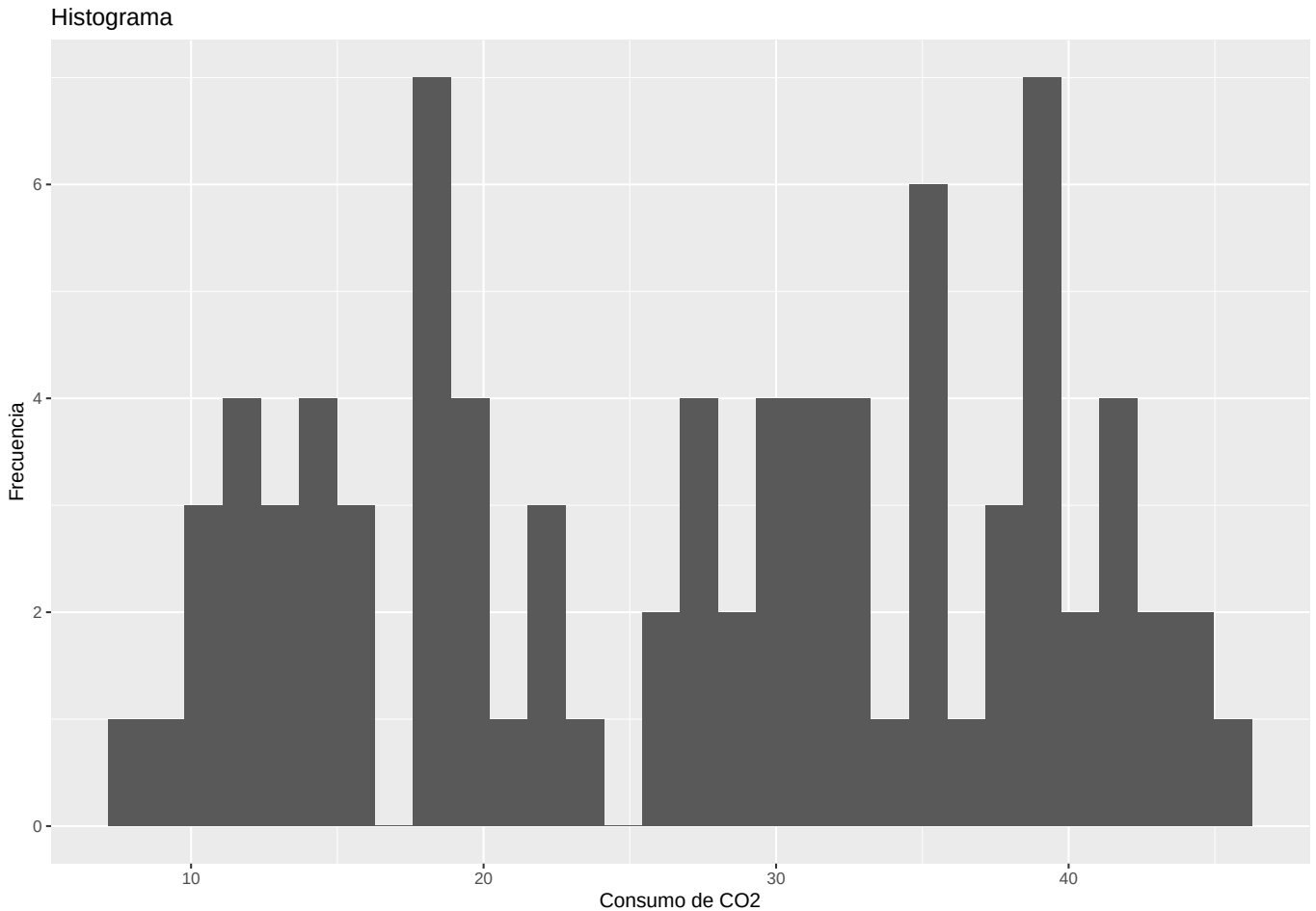
ggplot2 funciona por capas

```
ggplot(CO2, aes(uptake))
```



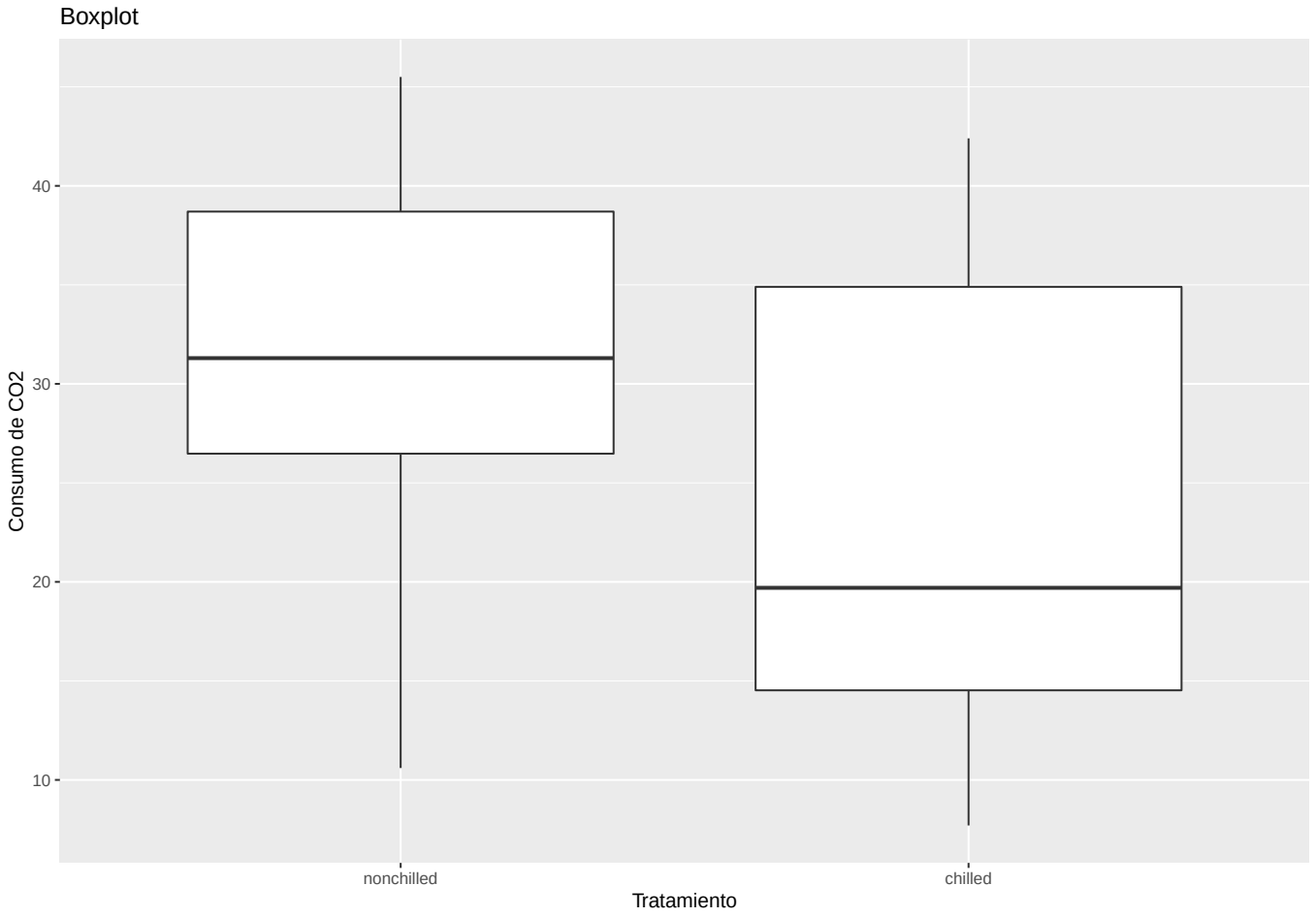
HISTOGRAMAS CON GGPLOT2

```
ggplot(CO2, aes(uptake))+
 geom_histogram()+
  labs(title="Histograma", x="Consumo de CO2",
       y="Frecuencia")
```



BOXPLOT CON GGPLOT2

```
ggplot(CO2, aes(x=Treatment, y=uptake))+
 geom_boxplot()+
  labs(title="Boxplot", x="Tratamiento",
       y="Consumo de CO2")
```



PRÁCTICA ANÁLISIS DE DATOS

1.- Guía de trabajo Rmarkdown disponible en drive.

2.- La tarea se realiza en Rstudio.cloud.

RESUMEN DE LA CLASE

- Identificamos variación, covariación, interacción y modelo que explica la relación entre variables.
- Realizamos gráficas avanzadas con ggplot2.
- Comunicamos EDA de forma efectiva.