CLASE 04 - ANÁLISIS EXPLORATORIO DE DATOS

Diplomado en Análisis de Datos y Modelamiento Predictivo con Aprendizaje Automático para la Acuicultura.

Dr. José Gallardo Matus

Pontificia Universidad Católica de Valparaíso

15 April 2023

PLAN DE LA CLASE

1.- Introducción

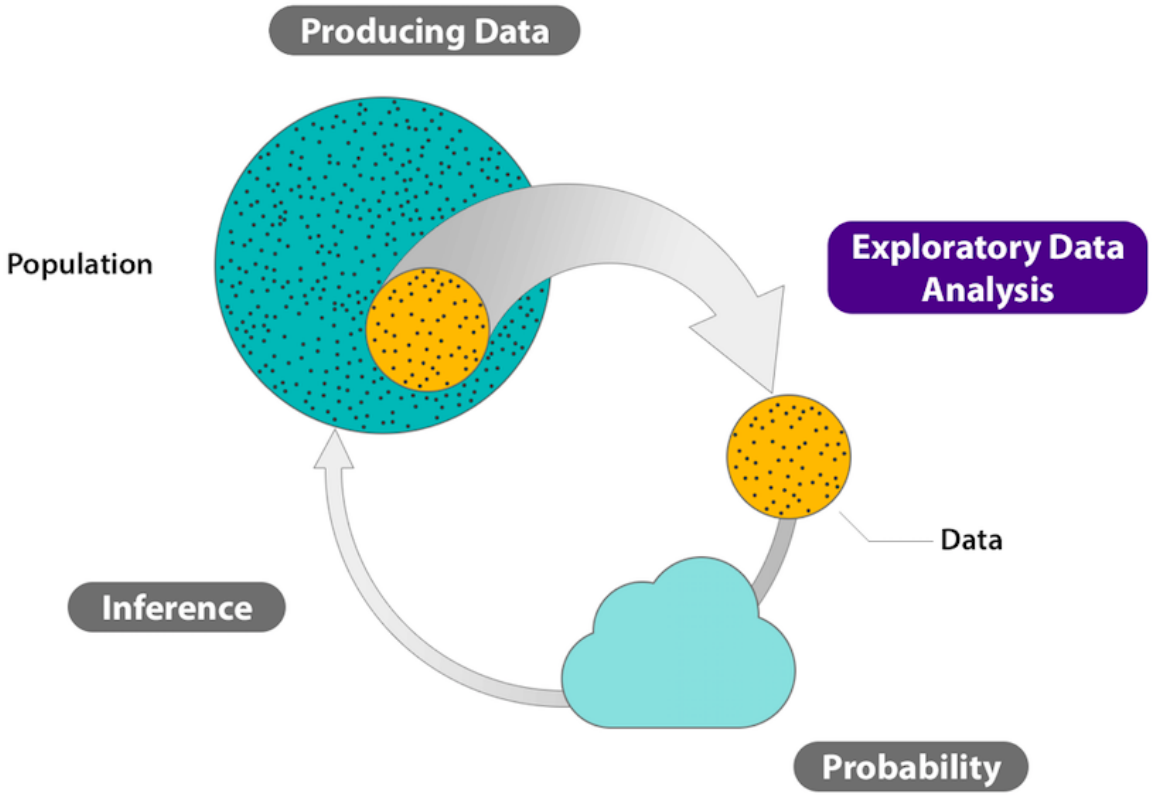
- ¿Qué es un análisis exploratorio de datos (EDA en inglés)?.
- ¿Por qué es importante?.
- Preguntas importantes para realizar un buen EDA.
- Gráficas con ggplot2.

2.- Práctica con R y Rstudio cloud

- 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 las variables/datos de un estudio.



¿POR QUÉ ES NECESARIO HACER UN EDA?

Principalmente para:

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

Diseño equilibrado o balanceado: Todos los tratamientos son asignados a un número equivalente de unidades experimentales (observaciones).

¿Datos son balanceados o desbalanceados?

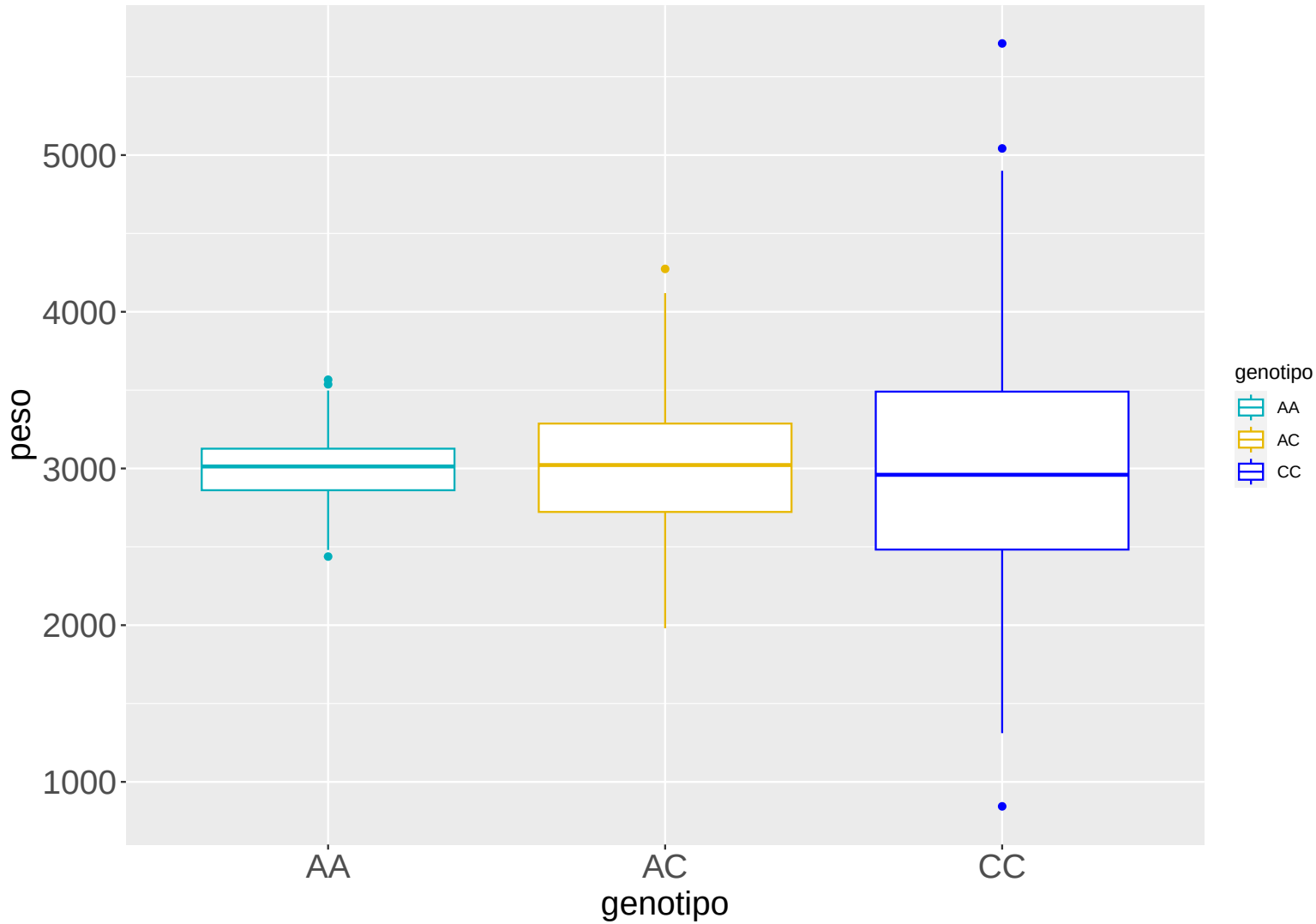
Tabla 1: Número de observaciones por sexo y dieta (D).

	D1	D2	D3	D4	D5	D6
Male	3	3	4	2	3	0
Female	9	7	8	9	11	12

Compare inferencia entre machos y hembras
Compare inferencia entre dieta 6 y otras dietas

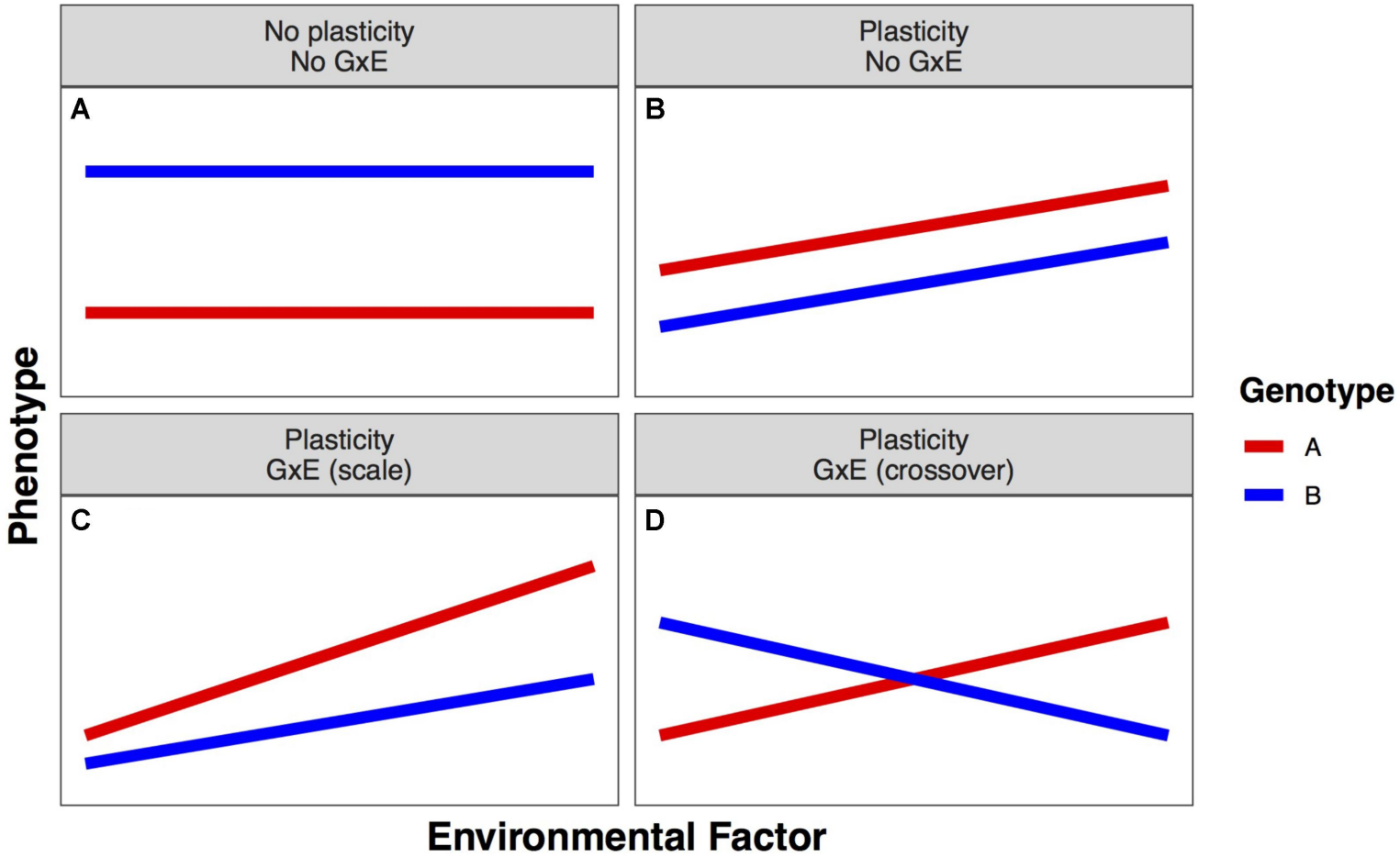
EDA: VARIACIÓN DENTRO DE UN FACTOR

¿La variación de mis datos es homogenea?



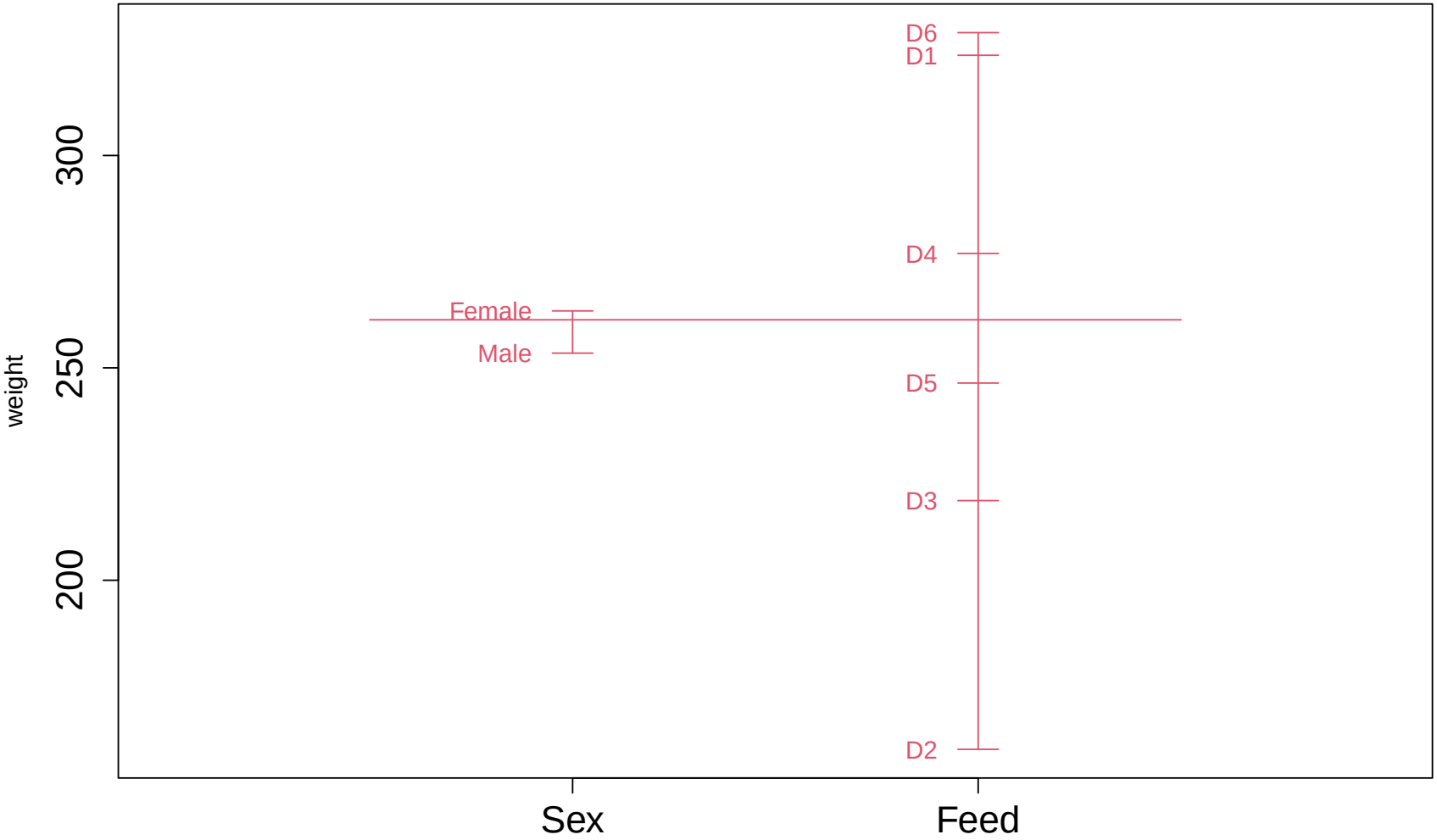
EDA: INTERACCIÓN ENTRE FACTORES

¿Existe interacción entre los factores?



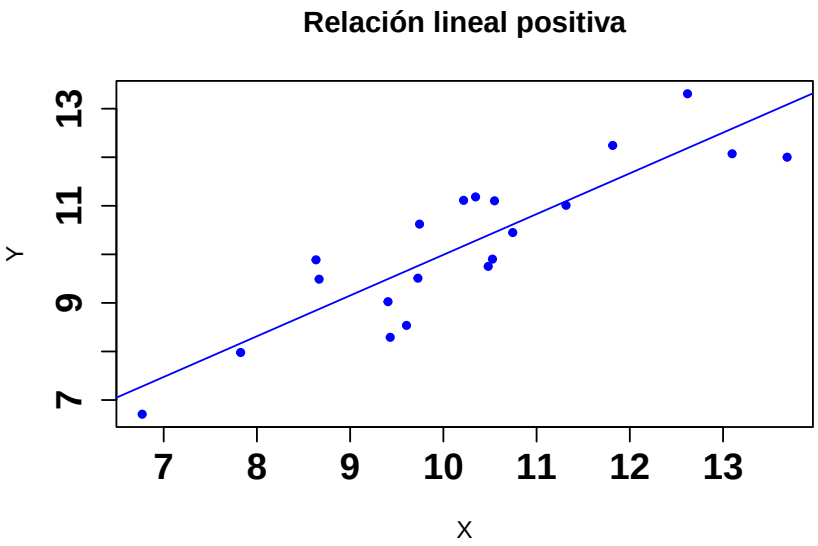
EDA: TAMAÑO DE LOS EFECTOS

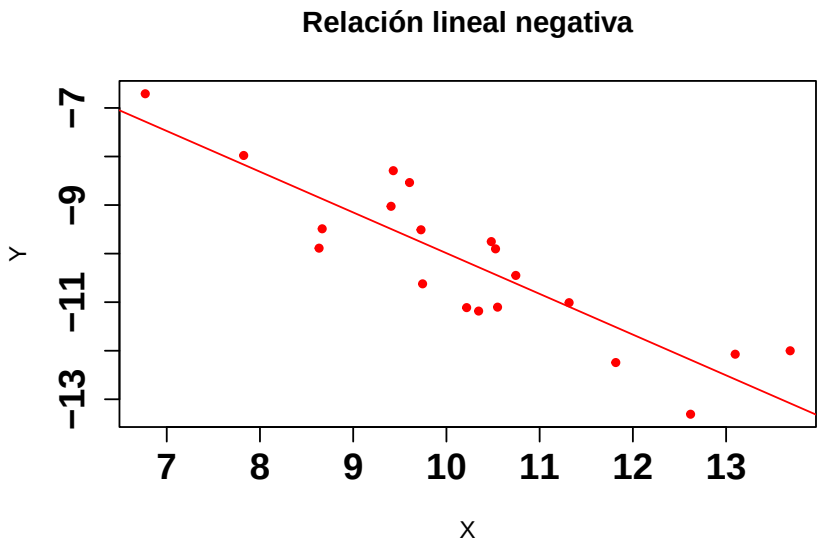
¿Qué factor tiene un mayor efecto sobre la variable respuesta?

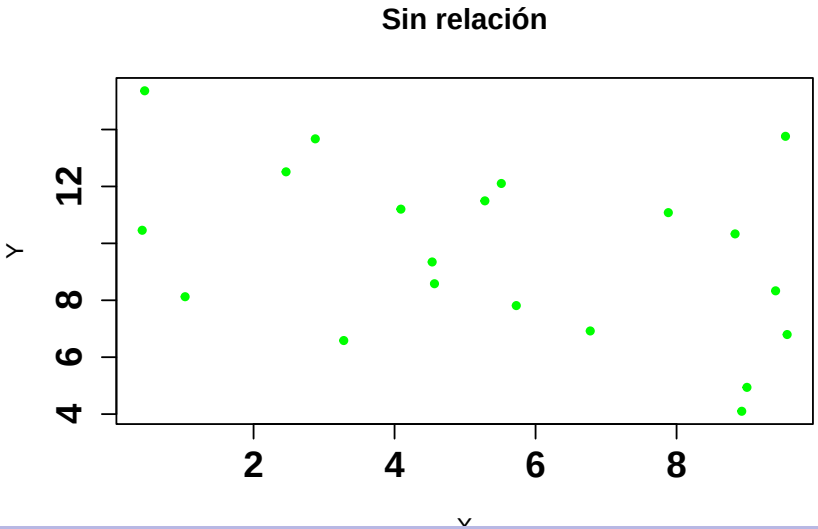


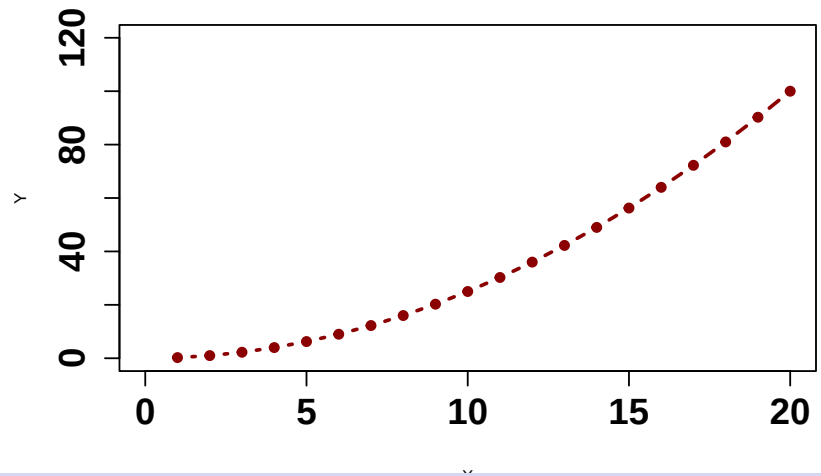
EDA: CORRELACIÓN

¿Existe covariación / correlación entre mis datos?









GRÁFICAS CON GGPLOT2

ggplot2
Paquete 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

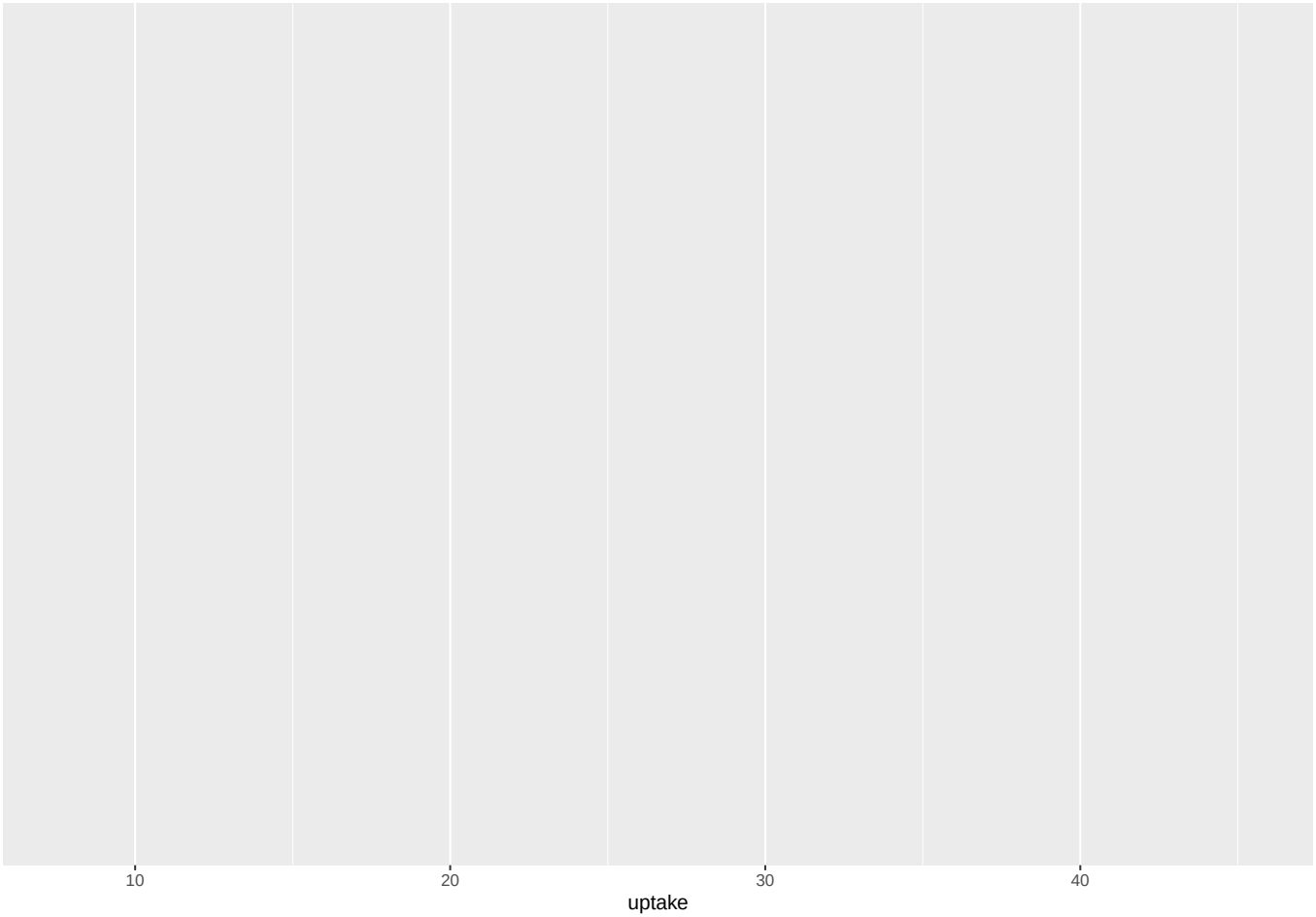
Comparación de algunos comandos de gráficas entre los paquetes **graphics** y **ggplot2**

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

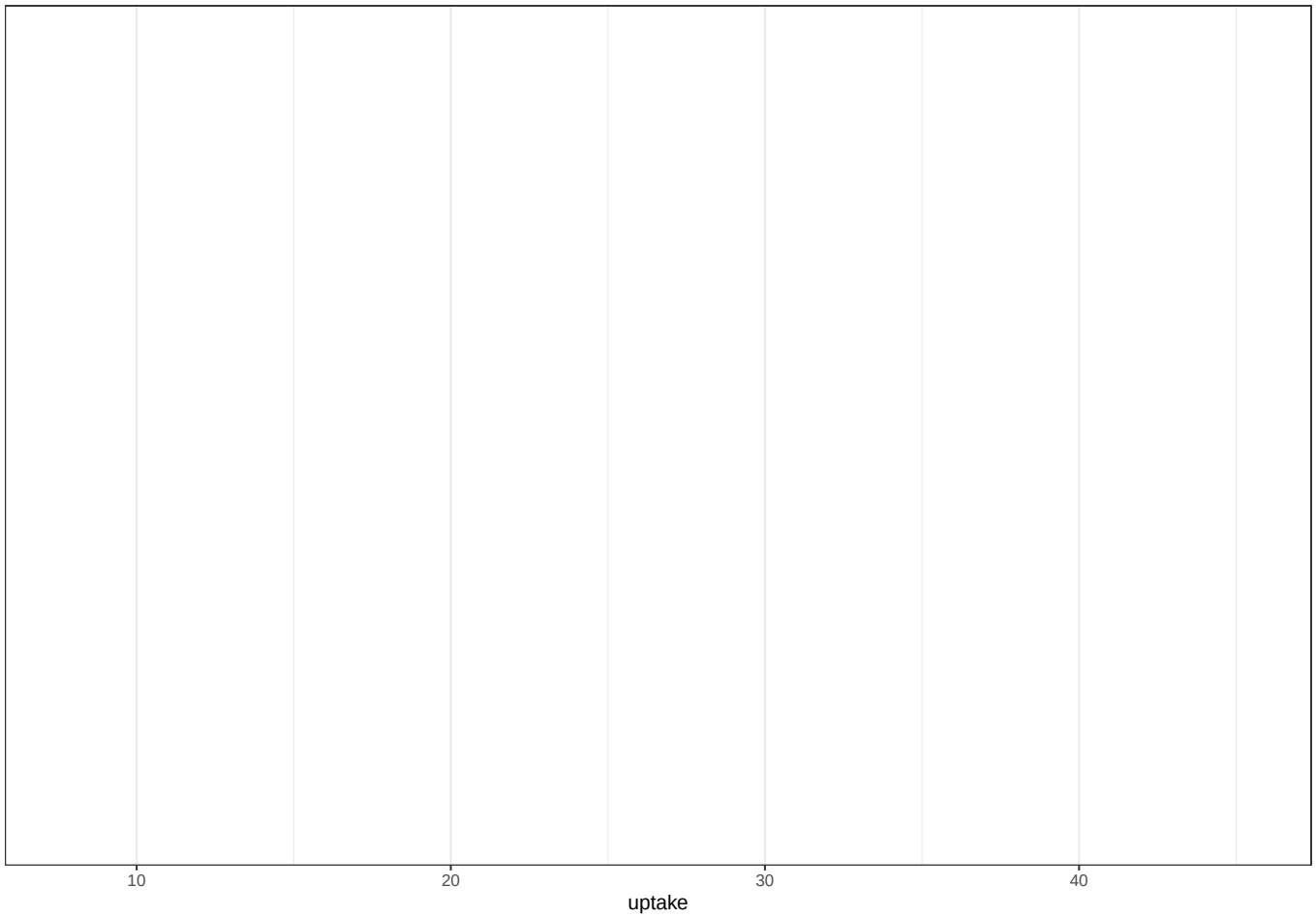
```
ggplot(CO2, aes(uptake))
```



TEMAS CON GGPLOT2

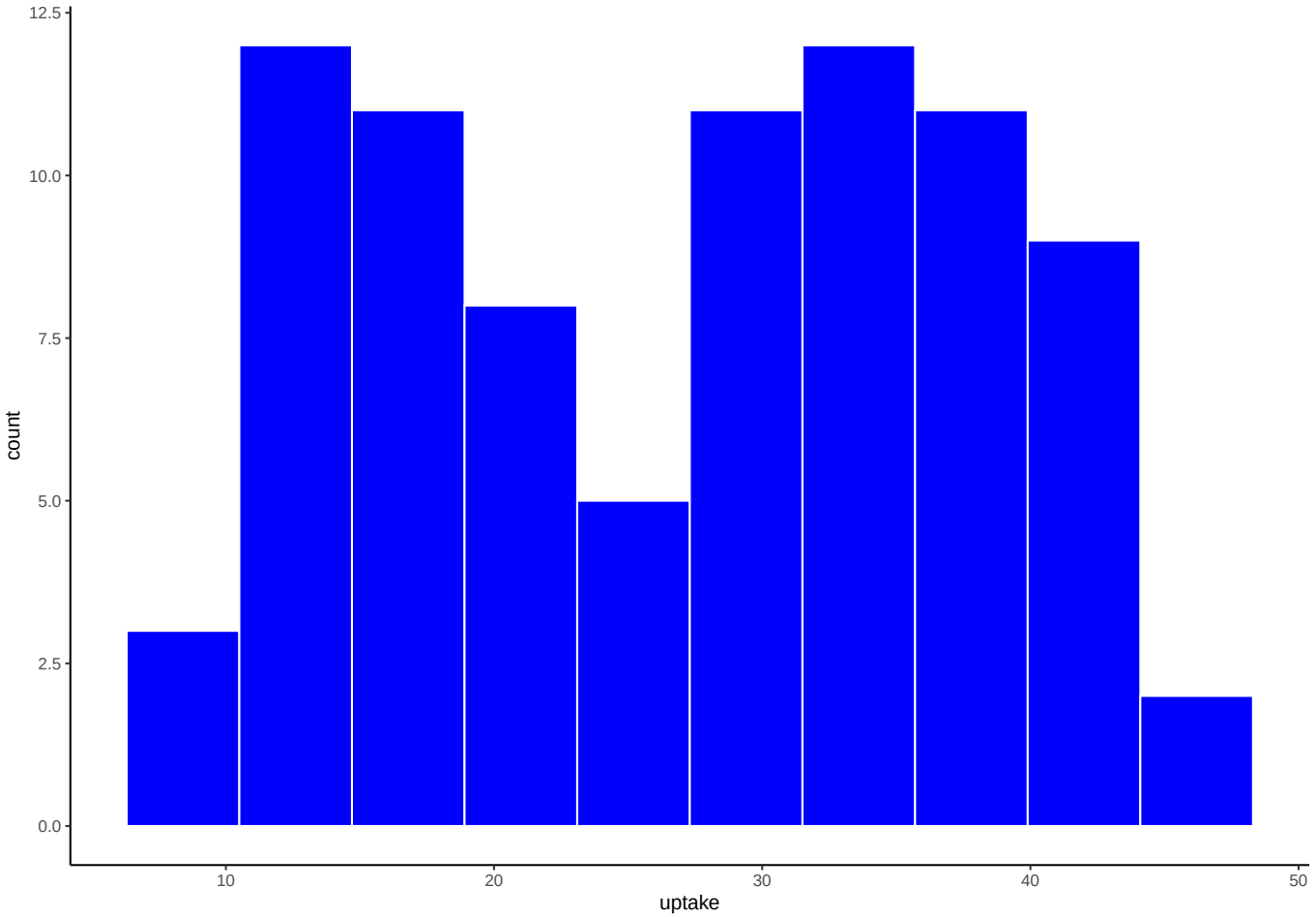
Diversidad de temas en ggplot2

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



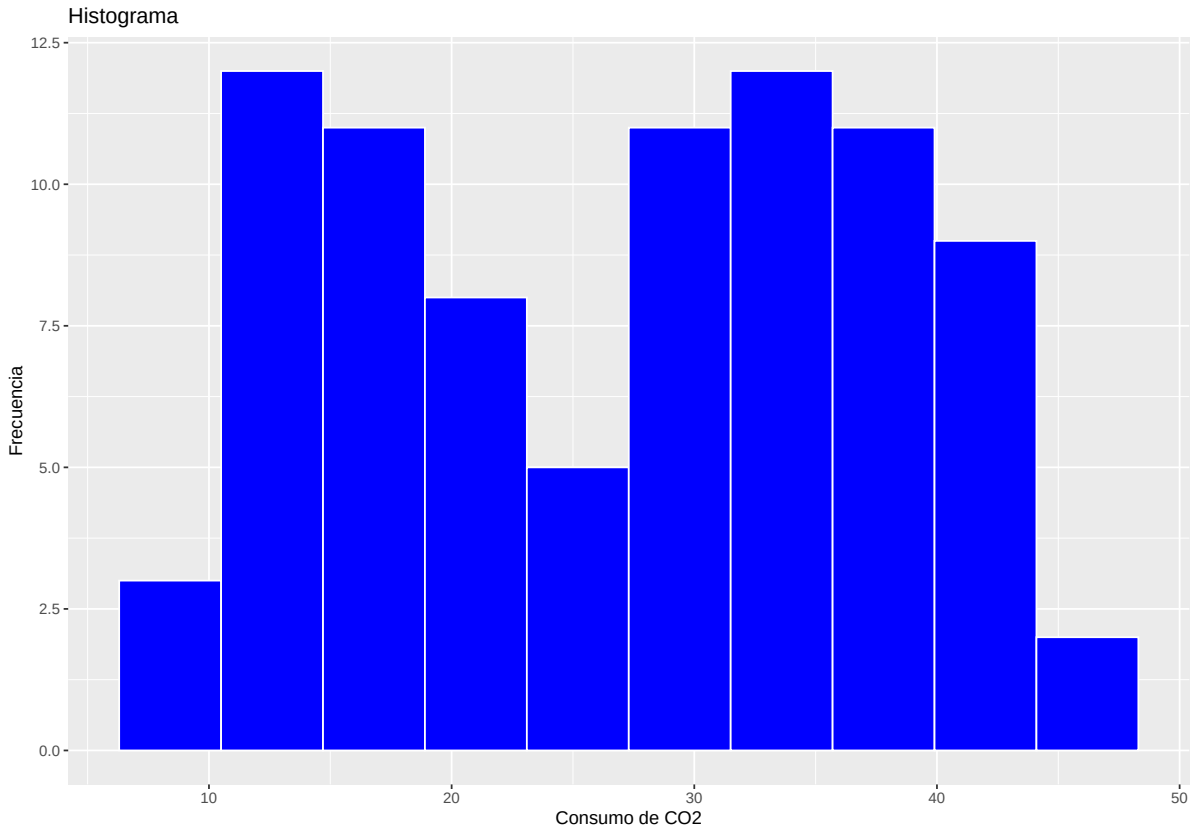
HISTOGRAMAS CON GGPLOT2

```
ggplot(CO2, aes(uptake)) +
  geom_histogram(color="white", fill="blue", bins = 10) +
  theme_classic()
```



CAMBIAR ETIQUETAS DE EJES

```
ggplot(CO2, aes(uptake))+
 geom_histogram(color="white", fill="blue", bins = 10)+
  labs(title="Histograma", x="Consumo de CO2",
       y="Frecuencia") + theme_gray()
```



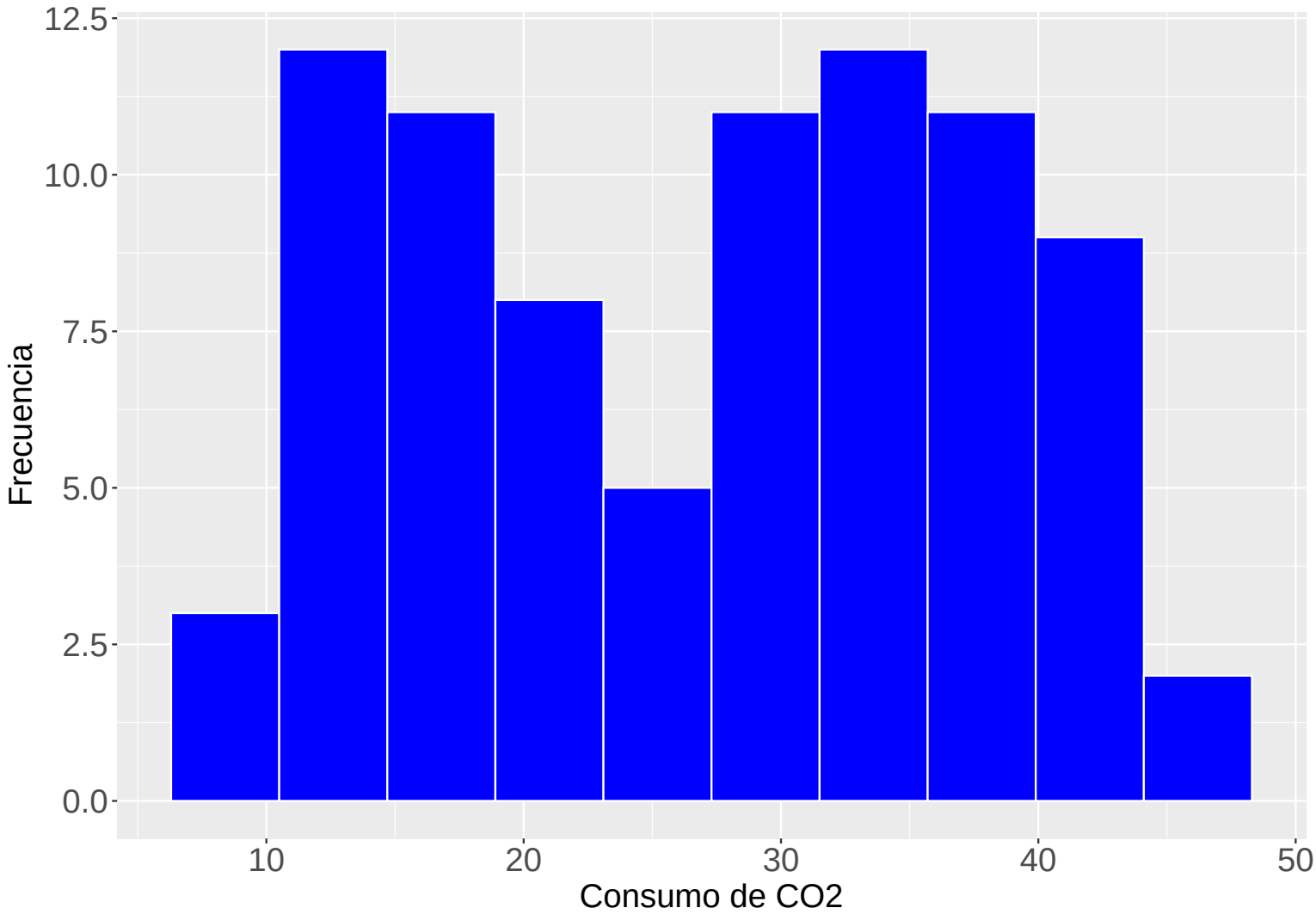
CAMBIAR TAMAÑO DE ETIQUETAS

```
My_Theme = theme(
  axis.title.x = element_text(size = 18),
  axis.text.x = element_text(size = 18),
  axis.title.y = element_text(size = 18),
  axis.text.y = element_text(size = 18))

plot_1 <- ggplot(CO2, aes(uptake))+
 geom_histogram(color="white", fill="blue", bins = 10)+
  labs(x="Consumo de CO2", y="Frecuencia")
```

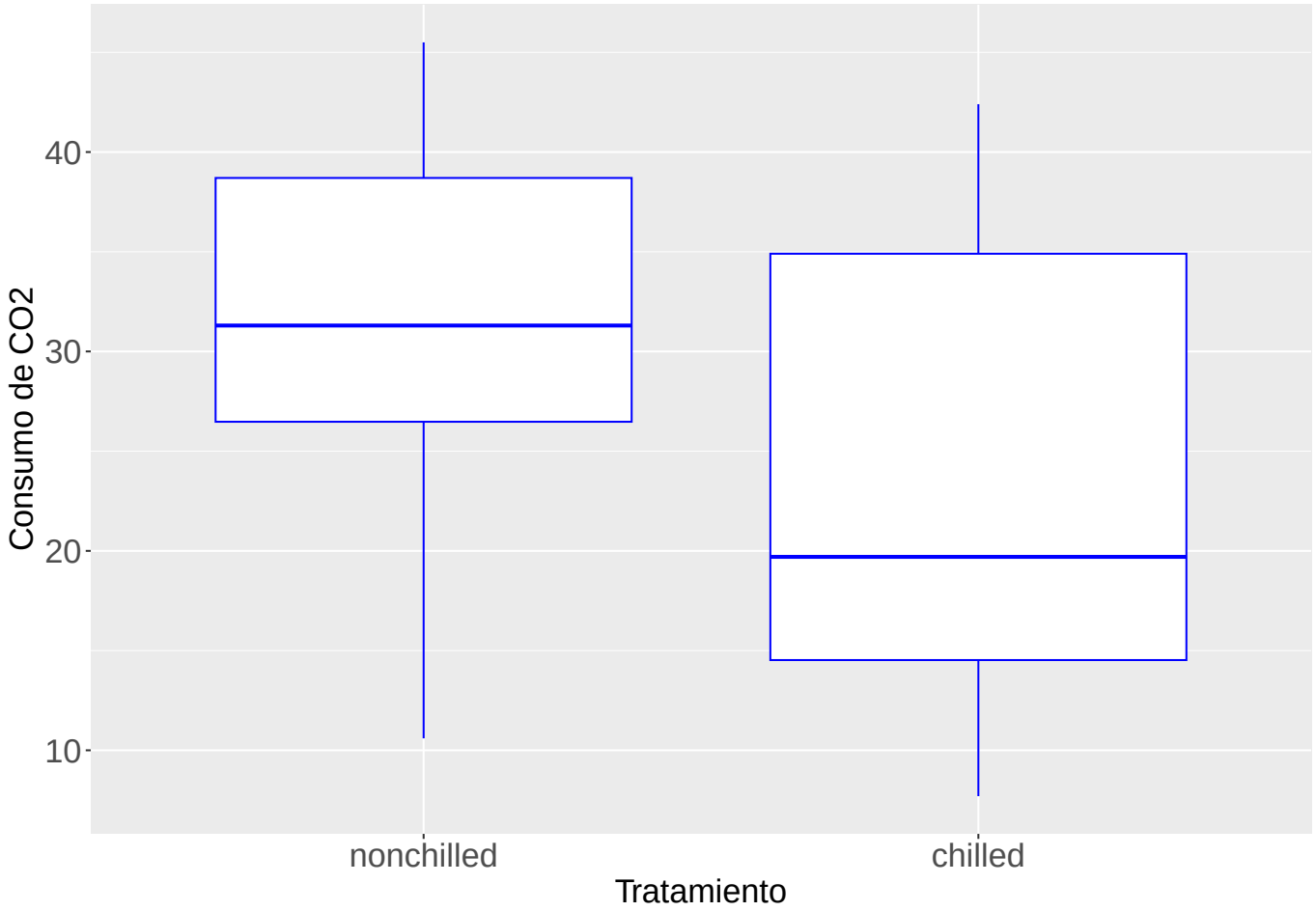
HISTOGRAMA FINAL

```
plot_1 + My_Theme
```



BOXPLOT CON GGPLOT2

```
ggplot(CO2, aes(x=Treatment, y=uptake))+
 geom_boxplot(color="blue")+
  labs( x="Tratamiento", y="Consumo de CO2") + My_Theme
```



RESUMEN DE LA CLASE

1. Importancia de los análisis exploratorio de datos.
2. Preguntas importantes de un EDA:
   2.1 Variación en mis variables de estudio.
   2.2 Covariación e interacción entre mis variables de estudio.
   2.3 Modelo más simple que explica la relación entre variables.
   2.4 Errores, datos faltantes, valores atípicos.
3. Realizamos gráficas avanzadas con ggplot2.
   3.1 Histograma y boxplot.