

# Clase 5 Análisis exploratorio de datos

Curso Introducción al Análisis de datos con R para la  
acuicultura.

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

09 July 2021

# PLAN DE LA CLASE

## 1.- Introducción

- ▶ Preguntas al curso y estudios de caso.
- ▶ ¿Qué es un análisis exploratorio de datos y por qué es importante?.
- ▶ Comunicar análisis exploratorios de forma efectiva.

## 2.- Práctica con R y Rstudio cloud

- ▶ Realizar un análisis exploratorio de datos.
- ▶ Realizar gráficas avanzadas con ggplot2.
- ▶ Elaborar un reporte dinámico en formato pdf.

# PREGUNTAS AL CURSO.

**\*\*Obj:\*\*** Identificar conceptos importantes para realizar un análisis exploratorio de datos.

- A. Varianza.
- B. Covarianza.
- C. Correlación.
- D. Interacción estadística.
- E. Variable respuesta o dependiente.
- F. Variable explicativa, predictora o independiente.

# ESTUDIOS DE CASO.

**Obj.** Identificar información relevante de un análisis exploratorio de datos.

¿Cuál es la variable respuesta?

¿Cuál es la variable explicativa o de clasificación?

¿Cuántos datos se muestran en la figura?

## **Casos de estudio**

Caso 1 - Informe sanitario salmonicultura 2019

Caso 2 - La Industria Camaronera de Ecuador 2020

Caso 3 - Toxicity of chemical substances in aquaculture 2019

Caso 4 - Respuesta inmune del camaron 2018

# 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?

# VISUALIZACIÓN DE DATOS AVANZADO CON GGPLOT2

**ggplot2:** Libreraría 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.



# EDA: IMPORTANCIA DE LA ESTRUCTURA DE LOS DATOS

**¿Mis datos son balanceados o no balanceados?**

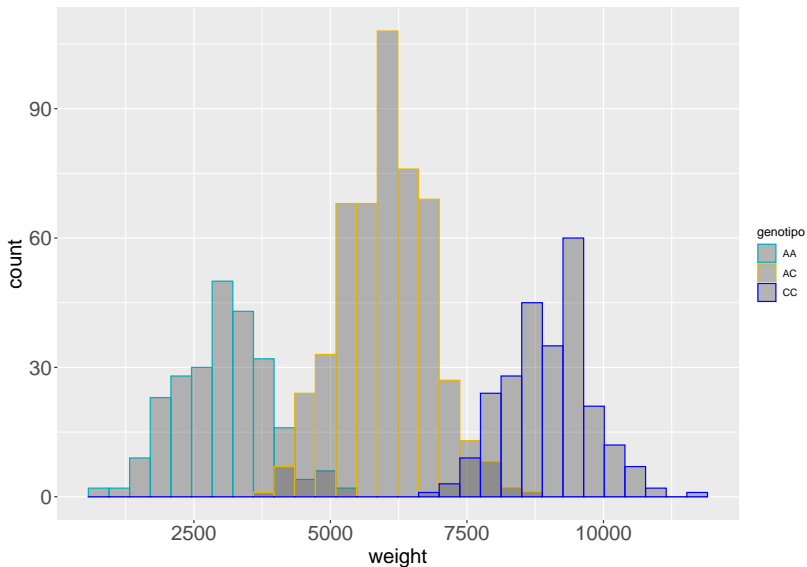
Table 1: Diseño no balanceado con datos faltantes

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



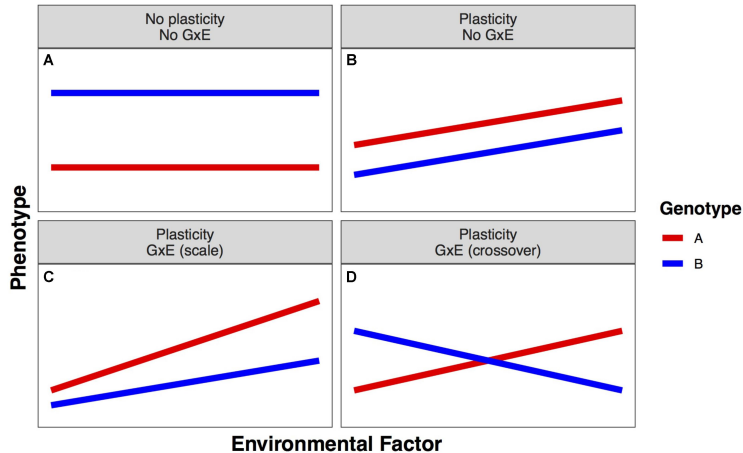
# EDA: VARIACIÓN DENTRO DE UN FACTOR

¿La variación de mis datos es homogénea?



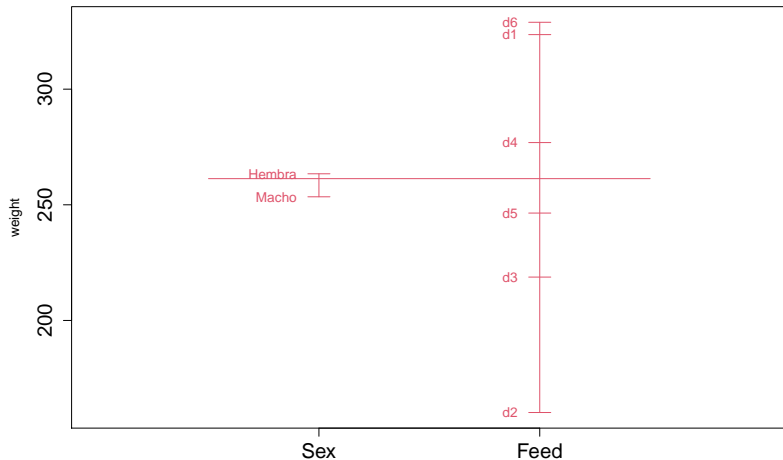
# EDA: INTERACCIÓN ENTRE FACTORES

¿Existe interacción entre los factores?



# EDA: TAMAÑO DE LOS EFECTOS

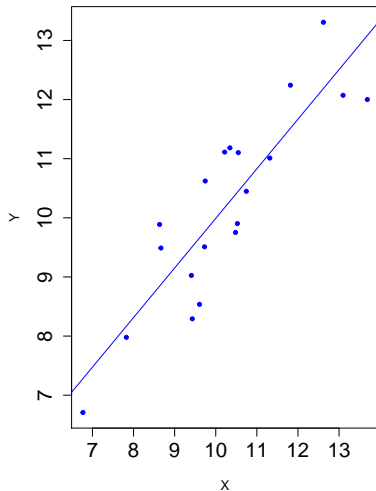
¿Qué factor tiene un mayor efecto sobre la variable respuesta?



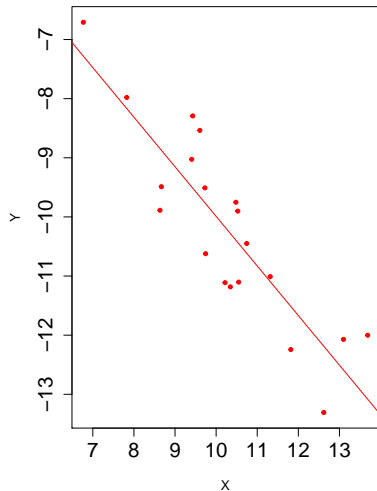
# EDA: CORRELACIÓN ENTRE VARIABLES CONTINUAS

¿Existe covariación / correlación entre mis datos?

Relación lineal positiva

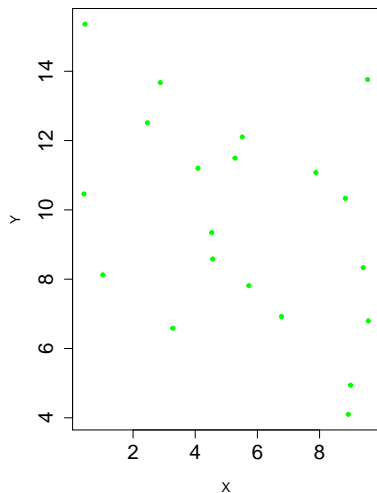


Relación lineal negativa

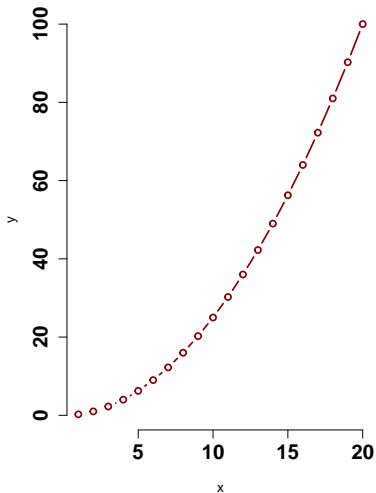


# EDA: OTRAS CORRELACIONES

Sin relación

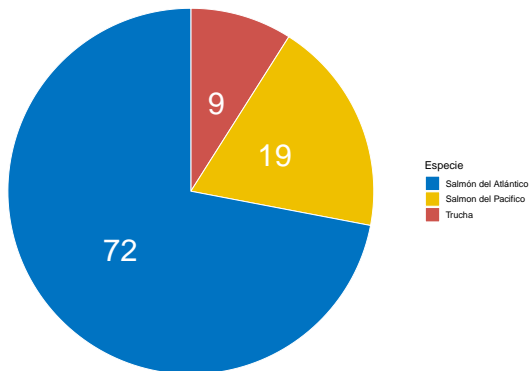


Relación no lineal



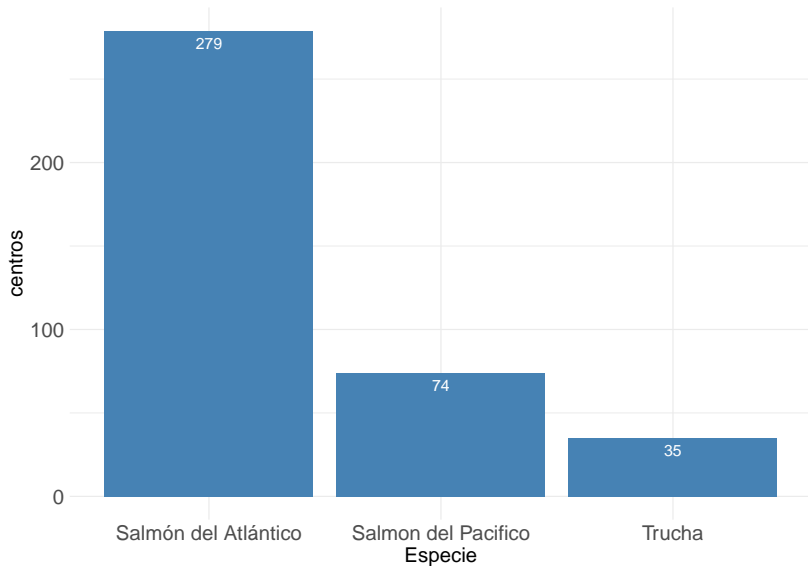
# COMUNICAR EDA DE FORMA EFECTIVA

## Evita las tortas



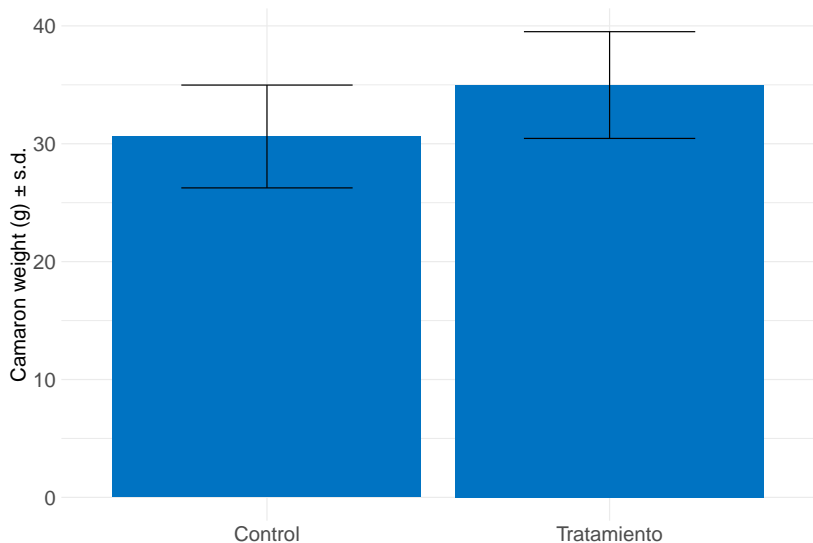
# SOLUCIÓN

## Prefiere datos brutos y barras



## COMUNICAR EDA DE FORMA EFECTIVA 2

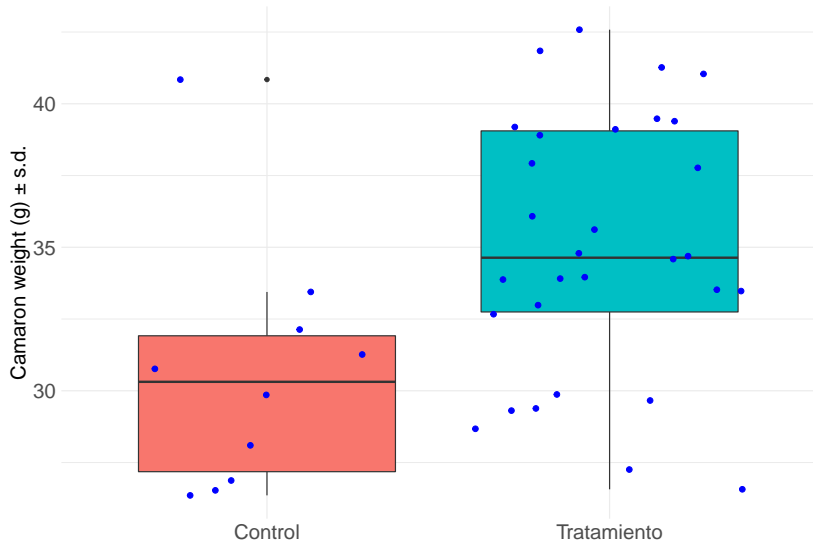
**Evita gráficas de barras para comparar grupos**





# SOLUCIÓN

**Prefiere boxplot, muestra tus datos  $\bar{x}$**



# PRÁCTICA ANÁLISIS DE DATOS

1.- Guía de trabajo Rmarkdown disponible en drive.

## **Clase\_5**

2.- La tarea se realiza en Rstudio.cloud. **05 Análisis exploratorio de datos**

# 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.