# Task 2

Tema: `ggplot2`
Profesor: Eduard F. Martínez González

Agosto 2025

## Introducción

El presente taller ha sido diseñado con el propósito de consolidar y aplicar los conocimientos adquiridos en torno a la limpieza y organización de datos, haciendo uso de la librería `dplyr` y de los contenidos desarrollados en la semana 3 del curso.

Para el desarrollo de las actividades propuestas, se pone a disposición una base de datos compuesta por 15.000 empresas de la ciudad de Cali. Es importante señalar que dicha base de datos fue construida a partir de información sintética, generada de manera aleatoria, y que no corresponde a datos reales. Su única finalidad es servir como insumo para los ejercicios académicos del presente taller. La base de datos incluye variables asociadas a aspectos relevantes de las empresas, tales como el número de sedes, el número de empleados, el volumen de ventas expresado en millones de pesos, el número de clientes y la duración de cada empresa durante los años 2024 y 2025.

## Repositorio

**Para resolver este taller**, el estudiante deberá **descargar y descomprimir** el repositorio `task-dplyr.zip`, disponible en el siguiente enlace: `aquí`. El repositorio contiene la siguiente estructura:

- Un archivo `proj.Rproj`, el cual debe ser utilizado para **iniciar RStudio**. Esto garantizará que el directorio de trabajo se establezca correctamente en la carpeta principal del repositorio.
- Una carpeta `code`, que incluye el **script base** sobre el cual deberá desarrollar la solución del taller. En caso de requerir librerías adicionales, estas deberán ser instaladas y llamadas al inicio del script.
- Una carpeta `input`, que contiene la **base de datos necesaria** para resolver los ejercicios propuestos.

Se enfatiza que el correcto uso del archivo `proj.Rproj` es indispensable para evitar errores

relacionados con rutas o directorios durante el desarrollo del taller.

## Instrucciones

- No seguir correctamente **todas** las instrucciones del taller implicará una penalización del **20 % sobre la nota total**.
- El taller puede desarrollarse en grupos de hasta dos personas. Únicamente uno de los integrantes debe cargar la entrega en la plataforma Intu.
- Este documento presenta dos opciones de trabajo —*Taller 1* y *Taller 2*—. **Cada grupo debe escoger únicamente uno de los dos talleres para su desarrollo.**
- La fecha límite de entrega es el **viernes 22 de agosto a las 10:00 a.m.**. A partir de ese momento la plataforma Intu se cerrará y no permitirá subir más archivos.
- La plataforma Intu recibirá exclusivamente dos archivos:
  - Un archivo en formato **.R** (script), con el desarrollo de los puntos del taller, documentado de manera clara y ordenada.
  - Un archivo en formato **PDF** (no Word ni ningún otro formato), que contenga las respuestas interpretativas solicitadas.
- El script en **R** debe cumplir con las siguientes condiciones:
  - Incluir al inicio: los nombres de los integrantes, la versión de R utilizada y la carga de todas las librerías necesarias.
  - Estar limpio y organizado: elimine funciones o líneas de código innecesarias.
  - Documentar claramente el desarrollo de cada punto utilizando comentarios: `## Punto 1`, `## Punto 2`, etc.

```r
## Nombre(s) de Autor(es)
## R version 4.5.0

## limpiar entorno
rm(list=ls())

## cargar librerias
require(dplyr)
require(skimr)
require(janitor)
require(rio)

## Punto 1
...
## Punto 2
...
```

# Opción 1: Fundamentos de Visualización con ggplot2

1. (15 pts) Cargue la base de datos y genere un gráfico de barras con el número de empresas por sector. Interprete qué sectores concentran más empresas.
2. (20 pts) Elabore un histograma del número de empleados. Interprete la forma de la distribución.
3. (20 pts) Genere un diagrama de dispersión entre `empleados` e `ingresos`. ¿Existe alguna relación aparente?
4. (15 pts) Cree un boxplot de ingresos por sector. Explique qué sectores muestran mayor dispersión.
5. (30 pts) A partir de los gráficos anteriores, redacte un breve texto (máximo media página) que resuma los hallazgos de manera narrativa.

# Opción 2: Storytelling con Personalización y Facetas

1. (15 pts) Cargue la base de datos y realice un gráfico de barras del número de empresas por sector. Personalice el gráfico con título, subtítulo, caption y un tema (`theme_minimal`, `theme_classic`, etc.).
2. (20 pts) Construya un diagrama de dispersión entre `empleados` e `ingresos`, diferenciando los puntos por color según el sector y tamaño según los ingresos.
3. (20 pts) Utilice `facet_wrap()` para mostrar la distribución de ingresos por comuna. Interprete al menos dos diferencias notables entre comunas.
4. (15 pts) Añada anotaciones: identifique las 5 empresas con mayores ingresos y etiquétenlas en el gráfico de dispersión.
5. (30 pts) Construya una mini-narrativa: seleccione una pregunta de negocio (por ejemplo, *¿dónde se concentran las empresas más grandes de Cali?*) y cree 2–3 gráficos que respondan a esta pregunta. Exporte uno de ellos en formato PNG con `ggsave()`.