Task 1

Tema: `dplyr`
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: Descriptiva y limpieza de datos

1. (30 pts) Cargue la base de datos en R y use la librería `skimr` para realizar una descripción general. Reporte:
   - Número de variables y observaciones.
   - Tipos de variables presentes.
2. (30 pts) Seleccione una variable numérica (por ejemplo, ventas o número de empleados) e interprete sus principales estadísticas:
   - Media, mediana y percentiles (25 y 75).
   - ¿Qué nos dice la distribución de esta variable acerca de las empresas?
3. (15 pts) Normalice los nombres de las variables usando la librería `janitor::clean_names` para eliminar espacios y mayúsculas.
4. (25 pts) Filtre la base de datos para conservar únicamente las observaciones con ventas positivas. Indique cuántas observaciones quedan en el nuevo objeto.

Opción 2: Descriptiva y manipulación con dplyr

1. (15 pts) Cargue la base de datos y normalice los nombres de las variables con `clean_names`.
2. (20 pts) Realice un análisis descriptivo con `skimr::skim` e interprete al menos una variable numérica en detalle (percentiles, media, distribución). Comente qué tan representativos son los valores centrales.
3. (15 pts) Filtre los datos para quedarse únicamente con las empresas que tienen ventas positivas y al menos 2 sedes. Indique cuántas observaciones cumplen estas condiciones.
4. (15 pts) Cree una nueva variable que calcule las ventas por empleado de cada empresa. Interprete el rango de valores que obtiene.
5. (20 pts) Usando `dplyr`, agrupe los datos por año y calcule:
   - Número promedio de clientes por empresa.
   - Ventas totales y promedio por año.
6. (15 pts) Genere una tabla resumen con las tres empresas con mayor número de empleados en 2025.