Taller 1 de Business Analytics

Tema: Limpieza de datos en R
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. Estos elementos permitirán desarrollar prácticas orientadas a la exploración, transformación y depuración de datos, reforzando las competencias en análisis aplicado.

Instrucciones

- No seguir correctamente **todas** las instrucciones del taller corresponde a una penalización del **20 % del total** de la nota.
- Este taller se puede desarrollar en grupos de hasta dos personas, y solo una persona del grupo debe subir el contenido a la plataforma Intu.
- Este documento presenta dos opciones de trabajo —*Taller 1* y *Taller 2*—, **los estudiantes deberán escoger un único taller para desarrollar en grupo**.
- La fecha máxima de entrega es el **viernes 22 de agosto a las 10:00 a.m.**. En ese momento se cerrará la plataforma Intu y no se permitirá cargar más archivos.
- La plataforma Intu solo permitirá cargar dos archivos: Un archivo en formato **.R** (script), en el que resolvieron los puntos del taller y otro archivo en formato **PDF** (no

Word ni ningún otro formato), que contenga las respuestas interpretativas solicitadas.

- Al inicio del script se debe: Mencionar la versión de **R** utilizada y se debe cargar todas las **librerías** necesarias para la resolución del taller. Tal como se muestra acontinuación:

```
## Nombre de Author
## R version 4.5.0

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

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

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

Taller 1: Descriptiva y limpieza de datos

Base 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.

Taller 2: Descriptiva y manipulación con dplyr

Base de datos:

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 accidentes 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.