Task 4

Tema: `Limpieza y exploración de datos`
Profesor: Eduard F. Martínez González

Septiembre 2025

Introducción

El presente taller ha sido diseñado con el propósito de consolidar y aplicar los conocimientos adquiridos en torno a la limpieza, exploración y organización de datos, haciendo uso de las librerías de `dplyr`, `janitor` y `skimr`, en coherencia con los contenidos desarrollados en la semana 7 del curso.

Para el desarrollo de las actividades propuestas, se pone a disposición una base de datos compuesta por 1.000 registros de pacientes atendidos en una clínica entre los años 2010 y 2025. 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 características sociodemográficas y clínicas de los pacientes, tales como edad, género, nivel educativo, afiliación a EPS, peso, estatura, fecha de registro y estado nutricional estimado mediante el índice de masa corporal (IMC).

Repositorio

Para resolver este taller, el estudiante deberá **descargar y descomprimir** el repositorio `task-tidy-data.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 para el grupo 007 es el **martes XX de septiembre a las 2:00 p.m.** y para el grupo 003 es el **viernes XX de septiembre 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 Taller 1: Limpieza, EDA y Reglas de Negocio

1. (20 pts) Cargue la base de datos y realice una **inspección inicial** usando funciones como `skim()` o `glimpse()`. ¿Qué problemas de calidad se observan en variables como `edad`, `nivel_educativo`, `eps`, `peso_kg` o `estatura_cm`?
2. (20 pts) Estandarice y documente los cambios en **nombres de columnas y tipos de variables** (ej. convertir `fecha_registro` a `Date`, asegurar que `edad` sea numérica). Justifique por qué estas transformaciones son necesarias.
3. (20 pts) Identifique y trate problemas de **valores perdidos (NA)** en variables clave. ¿El mecanismo parece MCAR, MAR o MNAR? ¿Qué estrategia aplicó (eliminar, imputar, modelar)?
4. (15 pts) Revise si existen **duplicados exactos o inconsistencias tipográficas** en variables categóricas como `eps` o `nivel_educativo`. ¿Cómo los resolvió?
5. (25 pts) Aplique **reglas de negocio** simples: verificar que `fecha_registro` esté entre 2010 y 2025, que `edad` sea positiva y que `imc` esté en un rango plausible (10–60). Resuma los principales hallazgos y explique cómo garantizaría trazabilidad y reproducibilidad en un pipeline real.

Total: 100 pts

Opción Taller 2: Transformaciones, Joins y Reportes

1. (15 pts) Cargue la base y reproduzca un pipeline de limpieza mínimo con `%>%`, incluyendo:
   - Normalización de nombres (`clean_names()`),
   - Conversión de tipos (`fecha_registro` a `Date`, numéricas como `edad`, `peso_kg`, `estatura_cm`),
   - Eliminación de duplicados.
2. (20 pts) Cree una variable de `imc` usando `peso_kg` y `estatura_cm`, y a partir de ella genere una variable categórica (`clasificacion_imc`) con `case_when()`. Resuma con `tabyl()` la distribución de categorías.
3. (20 pts) Simule una tabla externa (ej. regiones asignadas a pacientes) y únala a la base usando `left_join()`. Explique cómo los *joins* ayudan en la analítica de negocio.
4. (15 pts) Construya un pipeline que filtre pacientes **mayores de 65 años con IMC mayor a 30**, seleccione variables clave (`id`, `edad`, `genero`, `eps`, `imc`) y exporte a `.csv`.

5. (30 pts) Con `group_by()`, `summarise()` y `pivot_wider()`, cree un reporte que muestre **promedio de edad e IMC por género y EPS**. Interprete brevemente los patrones encontrados.

Total: 100 pts