Task 4

Tema: `Tidy-data`
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 y organización de datos, haciendo uso de la librerías de `dplyr` y `janitor`, y de 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.

```
## 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: Fundamentos de Limpieza y Exploración

1. (15 pts) Cargue la base de datos y use la función `skim()` para explorar el contenido. ¿Qué variables tienen más valores faltantes? ¿Cuáles presentan datos extremos?
2. (20 pts) Calcule el número de registros con valores incoherentes: edades negativas o mayores a 100, pesos menores a 10 kg o mayores a 250 kg, estaturas menores a 100 cm o mayores a 230 cm.
3. (20 pts) Cree una nueva variable llamada `imc` (índice de masa corporal) y clasifique a los pacientes según las categorías estándar (bajo peso, normal, sobrepeso, obesidad).
4. (15 pts) Elimine o corrija los valores atípicos en peso y estatura. Argumente su decisión.
5. (30 pts) Redacte un texto breve (máximo media página) explicando los principales retos de limpieza encontrados en la base de datos, qué decisiones se tomaron y por qué.

Total: 100 pts

Opción Taller 2: Perfilamiento de Datos y Validación de Calidad

1. (15 pts) Cargue la base de datos y use la función `janitor::tabyl()` para obtener el número de personas por género y nivel educativo. ¿Hay categorías mal escritas o no válidas?
2. (20 pts) Valide la coherencia entre edad y nivel educativo (por ejemplo, personas menores de 10 años con posgrado). Identifique cuántos casos presentan inconsistencias.
3. (20 pts) Analice la variable `fecha_registro`: ¿hay registros fuera del rango esperado (antes de 2010 o después de 2025)? Corrija o elimine esos casos.
4. (15 pts) Cree una tabla que muestre el promedio de edad y el IMC promedio por EPS. ¿Qué EPS agrupan a los pacientes con mayores valores?
5. (30 pts) Plantee una hipótesis de interés (por ejemplo: *¿existen diferencias en el IMC promedio por género?*) y explore la respuesta con código y gráficos. Guarde uno de los gráficos en PNG con `ggsave()`.

Total: 100 pts