Ejercicios Tutoría 3

El objetivo de esta tutoría es el de reforzar contenidos básicos pasados en el curso, **manejo y ordenamiento de bases de datos** con `dplyr` y `tidyr`, dos librerías de R que reúnen las funciones más eficientes para tal labor.

Parte 1: Limpieza de bases de datos y manejo de variables

En esta oportunidad usted trabajará con un extracto de la base de datos de la Encuesta Nacional de Salud de Chile realizada el año 2016. A continuación, presentamos sus variables y soporte de valores:

Variable	Etiqueta	Tipo	Soporte de valores
Zona	Macrozona geográfica	Categórica	"Norte", "Metropolitana", "Centro-Sur" y "Sur"
Comuna	Comuna de residencia	Categórica	"Arica", ..., "Natales"
Edad	Edad en años	Continua	15 – 98
Sexo	Sexo	Categórica	1= Hombre, 2=Mujer
Depresión	Padece depresión	Categórica	1=Si, 2=No, "Sin datos"
Fuma	Fuma cigarrillos	Categórica	1=Si, 2=No, "Sin datos"
Diabetes	Padece de diabetes	Categórica	1=Si, 2=No, "Sin datos"
presión_PAS	Presión arterial sistólica	Numérica	82.333 – 234.333
presión_PAD	Presión arterial diastólica	Numérica	43 – 131
Peso	Peso en kgs	Categórica	1=Si, 2=No, "Sin datos"
Talla	Talla (altura) en cms	Categórica	1=Si, 2=No, "Sin datos"
Asma	Padece de asma	Categórica	1=Si, 2=No, "Sin datos"

a. **Importación eficiente de los datos.** Importe la base de datos `ENS.RData` con la función correcta ¿por qué el formato requiere este método de importación? (*Hint: el formato RData es un formato nativo de R, que a diferencia de otras bases de datos, solo requiere de la función "load", y el objeto es guardado en el enviroment.*)

b. **Renombrar variables.** Los nombres de sus columnas tienen algunos nombres extraños. Cambie el nombre de estas variables de su nombre original al estilo `snake_case`. Luego, asigne etiquetas a sus variables.

c. **Manejo de variables en R.** Use la función `glimpse()` para observar la base de datos. Como puede observar, el tipo y clase de las variables presentan categorías equivocadas. En particular, se observan dos errores que deben ser arreglados:

i. **Tipo de dato de las variables.** Como puede observar, las variables `Depresión`, `Diabetes` y `Asma`, son reconocidos por R como `character` (*string*), pero sus valores son numéricos. ¿Es posible identificar la causa de este cambio de tipología? Cambie las variables al formato correcto.

ii. **Clase de las variables.** ¿Qué otros cambios debe realizar a sus variables respecto a su clase? Realice los cambios correspondientes a sus variables según la medición correspondiente. Al finalizar, exporte la base de datos en los formatos `RData`, `excel`, `stat` y `spss`.

Parte 2: manejo de base de datos con dplyr

a. **Creación de variables con `mutate`.** Genere las siguientes variables continuas y categóricas y añádalas a la base de datos:

   i. `hta`: **Diagnóstico de Hipertensión.** Utilice las varialbes de `presion_pad` y `presion_pas` para construir la variable "padece hipertensión" (`hta`). Según la OMS, se considera que el paciente tiene la presión alta cuando la presión está por sobre los 140/90 mmHg (PAS/PAD)[1].

   ii. `grupos_edad`: **grupos de categoría de la edad.** Agrupe las siguientes edades en sus categorías de edad: `"15 a 24 años"`, `"25 a 44 años"`, `"45 a 64 años"` y `"65 años o más"`.

   iii. `indice_riesgo`: **Índice de Factores de Riesgo.** En una base de datos aparte, genere un índice en donde se suman la cantidad de afecciones a la salud presentes en la base de datos.

b. **Generación y visualización de tablas con kable.** Genere las siguientes tablas en formato académico. Asegúrese de facilitar el formato correcto apoyándose en las herramientas que R brinda para tal ocasión.

   i. Una tabla con los valores perdidos de las variables.
   ii. Una tabla de estadísticos descriptivos de las variables continuas.

c. **Funciones de la familia `join`.** Las bases de datos `opiniones_1.xlsx` y `opiniones_2.xlsx` contienen diferentes frases de pacientes con respecto a sus síntomas. Tanto en la primera base de datos como en la segunda base de datos algunos pacientes se repiten.

   i. Realice dos uniones de esta base de datos: una que involucre a todos los pacientes, y otra que involucre solo a los pacientes que se encuentren en ambas bases de datos.

[1] Puede comprobar la fuente en: https://www.paho.org/es/temas/hipertension#:~:text=La%20presi%C3%B3n%20arterial%20alta%20igual,para%20muertes%20por%20enfermedades%20cardiovasculares.

ii. Arme una tabla de datos que muestre el conteo de las palabras de la base de datos unificada. Una vez obtenida la tabla, use el método más eficiente para eliminar los conectores.

Parte 3: introducción al manejo de estructura de los datos con `tidyr`

a. **Pivotear una tabla de datos.** Una dos bases de datos `pacientes_1.xlsx` y `pacientes_2.xlsx` para ello, uilice el mejor método. Exporte una base de datos en formato *wide* y otra base de datos en formato *long*.