**Tarea 2: Importación, limpieza e identificación de errores y valores atípicos en R**

**Instrucciones:** **Instrucciones:** Resuelve los siguientes problemas utilizando R y contesta el cuestionario anexado en Classroom. **IMPORTANTE:** Debes subir el script con las respuestas. Asegúrate de que tu código esté bien comentado y organizado.

**Fecha Límite de Envío: 31 de octubre de 2024**

**PARTE 1: Dataset Cancer**

**1. Carga y Exploración Inicial del Dataset**

1. Carga el dataset cancer.
2. Visualiza las primeras 10 filas del dataset con la función head().
3. Describe brevemente la estructura de los datos utilizando la función str().
4. **¿Cuál es el valor de "Tiempo de Sobrevida" del cuarto paciente?**
5. **¿Cuál es el número de filas y columnas que contiene el data.frame de "cancer"?**

**2. Limpieza de Nombres de Variables**

1. Utiliza la función clean\_names() del paquete janitor para estandarizar los nombres de las variables en cancer.
2. Asigna el resultado a un nuevo objeto llamado cancer\_clean.
3. Imprime los nombres de las variables antes y después de la limpieza.
4. Contesta las pregunta en el cuestionario de Classroom.
5. **Suba un screenshot del nombre nuevo de las variables**

**3. Introducción de Valores Faltantes y su Identificación**

1. Identifica cuántos valores faltantes (NA) hay en cada columna del dataset cancer\_clean.
2. **¿Cuál es la variable que tiene la mayor cantidad de variables perdidas?**

**4. Filtrado de Filas con Valores Faltantes**

1. Filtra las filas de cancer\_clean que tienen al menos un valor faltante y guárdalas en un nuevo dataset llamado cancer\_clean\_complete.
2. **¿Cuánto es el número de filas que resta posterior al filtrado de las variables?**

**PARTE 2: Dataset framinham.csv**

**Paso 1: Selección de Variables**

Selecciona las siguientes variables en un nuevo dataset:

* male
* age
* diabetes
* glucose
* sysBP
* diaBP
* BMI

**Paso 2: Recodificación de Nombres de Variables**

Recodifica el nombre de todas las variables al español:

**“Variable original -> variable nueva”**

* male -> sexo
* age -> edad
* diabetes -> diabetes
* glucose -> glucosa
* sysBP -> TAS
* diaBP -> TAD
* BMI -> IMC

**Paso 3: Creación de Variables Categóricas**

Crea las siguientes variables categóricas:

1. **Categorías de Edades:**
   * Grupo 1: Edad ≤ 45
   * Grupo 2: Edad 45 a 65
   * Grupo 3: Edad ≥ 65 años
2. **Diabetes:**
   * 1: Antecedente de diabetes o glucosa mayor a 128 mg/dl
   * 0: No cumple con las condiciones anteriores
3. **Hipertensión Arterial:**
   * 1: Antecedente de hipertensión o TAS ≥ 140 mg/dl o TAD ≥ 90 mg/dl o uso de medicamentos
   * 0: No cumple con las condiciones anteriores
4. **Categorías de IMC:**
   * Infrapeso: IMC < 18.5
   * Normopeso: IMC 18.5 a 24.9
   * Sobrepeso: IMC 25 a 30
   * Obesidad: IMC ≥ 30

**Paso 4: Conversión a Factor**

Convierte estas nuevas variables a factores y asigna las etiquetas correspondientes según el inciso anterior.

**Responde las Siguientes Preguntas:**

1. **¿Cuál es la frecuencia de hombres y mujeres?**
2. **¿Cuál es la frecuencia de grupos de edades?**
3. **¿Cuál es la frecuencia de diabetes presente?**
4. **¿Cuál es la frecuencia de hipertensión arterial presente?**
5. **¿Cuál es la frecuencia de grupos de IMC?**
6. **Suba un screenshot de la evidencia de su codificacion. Puede usar la siguiente función:**
   1. **“base %>% summary()”**

**Punto Extra:**

1. ¿Cuál es el promedio y desviación estándar de la edad entre hombres y mujeres?
2. ¿Cuál es la frecuencia de diabetes entre hombres y mujeres?