Introducción al análisis de datos con R

Introducción a la limpieza y manipulación de datos

Ejercicios - Sesión 2

5 - 9 junio de 2023

1. Introducción

En este ejercicio, vamos a preparar los datos de la Encuesta de Educación y Hogares de Andalucía del año 2010. Esta encuesta proporciona información sobre los hijos, padres y hogares. Puedes obtener más información en el siguiente enlace: https://www.juntadeandalucia.es/institutodeestadisticaycartografia/descarga/encSocial/2010/encSocial2010.htm

El objetivo de este ejercicio es limpiar una base de datos y prepararla para su posterior análisis. Para ello, vamos a obtener una base de datos que incluya las siguientes variables:

- Rendimiento educativo: puntuación en el test de matemáticas.
- Variables sociodemográficas básicas: sexo del adolescente y cohorte de nacimiento.
- Variables de contexto socioeconómico: nivel educativo de los padres.
- Una variable de pesos poblacionales.

2. Primeros pasos

- 2.1. Descarga los datos y guárdalos en tu escritorio.
- 2.2. Abre un nuevo script en R y limpia el espacio de datos ejecutando `rm(list = ls())`.
- 2.3. Establece el directorio de trabajo en la carpeta donde se ubican los datos.
- 2.4. Instala y carga las librerías necesarias para el análisis.
- 2.5. Importa el archivo de datos llamado `ESOC` que se encuentran en formato csv y asígnalo a un objeto llamado `esoc.`

  Consejo: durante el proceso de manipulación de datos, es aconsejable tener abierto y echar un vistazo al documento "Diseño de registro" que muestra toda la información de las variables recogidas en la encuesta.

3. Limpiando los datos

- 3.1. Abre el archivo de "Diseño de registro" y observa el nombre, descripción y categorías de las variables.
- 3.2. Selecciona con `select()` las variables y crea a un nuevo data frame llamado **df_esoc**.

- 3.3. Usando el data frame **df_esoc**, recodifica las variables usando `mutate()` o `transmute()`:
  - 3.3.1. Renombra la variable SEXO_EGO a "sexo". Utilizando `case_when()`, recodifica el valor 1 como "chico" y el valor 6 como "chica". Trata esta variable como un factor utilizando `as.factor()`.
  - 3.3.2. Renombra la variable SUBP a "cohorte". Utilizando `case_when()`, recodifica el valor 1 como "1994" y el valor 2 como "1998". Trata esta variable como un factor utilizando `as.factor()`.
  - 3.3.3. Renombra la variable STUDIOSC a "estudios_padres". Utilizando `case_when()`, recodifica los valores 1, 2 y 3 como "Secundaria básica o menor", los valores 4 y 5 como "Secundaria superior", y el valor 6 como "Universitarios". Trata esta variable como un factor utilizando `as.factor()`.
  - 3.3.4. Renombra la variable RMATE a "test_mates". Elimina los valores "No procede" o "No consta" utilizando esta línea de código: `if_else(RMATE == -1, NA, RMATE)`. Realiza un resumen de la variable utilizando `summary()` para observar sus valores mínimos, máximos y la media. Trata esta variable como numérica utilizando `as.numeric()`.
  - 3.3.5. Renombra la variable FEIR a "pesos". Trata esta variable como numérica utilizando `as.numeric()`.

4. Análisis descriptivo de los datos

- 4.1. Observa si existen valores perdidos en cada variable del conjunto de datos. Utiliza la siguiente línea de código para hacerlo: `df_esoc %>% summarise_all(~sum(is.na(.)))`. Si existen, elimina todos los casos con datos faltantes utilizando `drop_na()`.

Calcula la media de calificación en los test de razonamiento matemático para los alumnos de primaria y secundaria. Para ello, ten en cuenta los siguientes pasos:

- 4.2. Crea dos data frames distintos para la cohorte de 1994 (llámalo df_secundaria) y 1998 (df_primaria). Utiliza `filter()` y los operadores lógicos necesarios para realizar esta operación.
- 4.3. Calcula la media para cada cohorte en cada data frame por separado, utilizando `summarise()` y `mean()`.
- 4.4. Calcula la media de calificaciones en los tests de matemáticas para chicos y chicas de secundaria. Para ello, utiliza `group_by()` y `summarise()`. Además, utiliza `weighted.mean()` para incluir los pesos poblacionales y asegurar que las estimaciones sean representativas de la población objeto de estudio.
- 4.5. Calcula la media de calificaciones en los tests de matemáticas para chicos y chicas de secundaria y primaria de diferentes orígenes socioeconómicos (según los estudios de los padres). Para ello, utiliza `group_by()` y `summarise()`. Además, utiliza `weighted.mean()` para incluir los pesos poblacionales y asegurar que las estimaciones sean representativas de la población objeto de estudio. Por último, utiliza `arrange()` para ordenar los valores del data frame.