Unidad 03:
Lectura de datos

Mg. Jesús Eduardo Gamboa Unsihuay

Octubre del 2024

Lectura de datos estructurados

Archivos de datos

Los datos pueden estar contenidos en archivos de distintos formatos, entre los que se tiene los más comunes:

1. Archivo de texto plano: su contenido es netamente textual, sin ningún formato. Lleva la extensión **.txt** y puede ser creado y/o editado en Bloc de notas.
2. Archivo de valores separados por comas: permite representar datos en formato tabular, en el que las columnas son separadas por comas (o puntos y comas, dependiendo del país y/o la configuración de la computadora). Lleva la extensión **.csv** y puede ser creado y/o editado en programas como Bloc de notas o Excel.
3. Archivo de hoja de cálculo: permite almacenar los datos en celdas, dispuestas en filas y columnas. La extensión común es **.xlsx** o **.xls**

Antes de iniciar con la lectura de datos...

... debemos entender y ordenar nuestro espacio de trabajo

Entorno de trabajo

¿Cómo RStudio accede al archivo que contiene nuestro conjunto de datos? ¿Dónde "**vive**" nuestro código de análisis de datos?

Ejecute el código:

```
getwd()
```

Por defecto, R ejecutará todos nuestros análisis en esta carpeta. ¿Cómo cambiamos esa ruta?

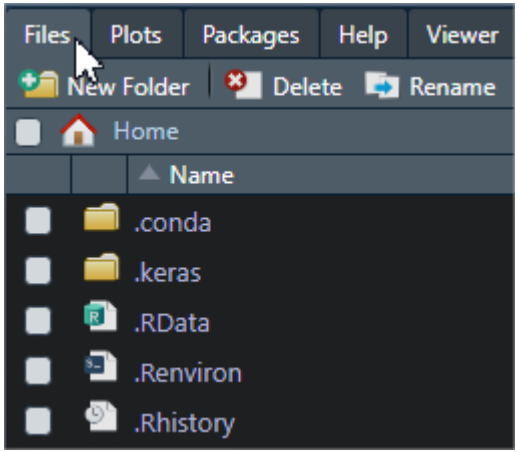
1. Usar la función `setwd(...)`
2. **Usar proyecto <- Reproducible**

Entorno de trabajo

¿Quiénes son los “**vecinos**” de nuestro código de análisis de datos?

Puedes verlo en la pestaña **Files** del panel inferior derecho.

Se listan todos los archivos que conforman la carpeta que obtuvimos con la función `getwd()` o la que fijamos con `setwd(...)`



Entorno de trabajo

La reproducibilidad es importante. Si tú ejecutas un código y se lo compartes a otra persona, debe obtener el mismo resultado, sin necesidad de realizar cambios en el código.

¿Qué sucede si un código comienza con `setwd("C:/Users/Usuario 1")`?

Proyectos en RStudio

Permite manejar el entorno de trabajo de manera ordenada. Todos los archivos quedan empaquetados en una carpeta de trabajo.

¿Cómo crear un proyecto en RStudio?

1. Crear una **carpeta de trabajo**, la cual contiene o contendrá todos los archivos necesarios para el análisis.
2. Click en File → New Project o hacer click en los íconos que ejecutan esa misma acción
3. Elegir “Existing Directory” y luego en “Browse” buscar la **carpeta de trabajo**

Lectura de datos

Utilizaremos las funciones:

```
read.table, read_tsv (" ")

read.csv, read_csv (",")

read.csv2, read_csv2 (";")

read.delim, read_tsv ("\t")
```

Necesitaremos los paquetes readr y readxl

`read.table(...)`

Argumento	Tipo	Descripción
file	Character Obligatorio	Archivo de datos a ser leído. Va entre comillas. Incluir su extensión (txt, csv), por ejemplos **`"datos336.txt"`**
header	Logical Opcional	`TRUE` si la primera línea contiene los nombres de las variables. Por defecto se asume que es **`FALSE`**, a menos que el número de campos en la primera fila es uno menos que el número de columnas
sep	Character Opcional	Caracter que indica el separador entre columnas. Va entre comillas, por defecto asume que es **`sep = " "`** (al menos un espacio en blanco).
dec	Character Opcional	Caracter que indica el separador de la parte entera y decimal en los números, por defecto asume **`dec = "."`**

read.table(...)

Argumento	Tipo	Descripción
col.names	Vector Opcional	Vector que indica el nombre de las columnas
row.names	Vector Opcional	Vector que indica el nombre de las filas. También se puede indicar que una columna contiene estos nombres.
nrows	Integer Opcional	Número máximo de filas a leer
skip	Integer Opcional	Número de filas al inicio del archivo que no serán leídas
na.strings	Character Opcional	Caracter que indica cómo se declaran los valores perdidos (Not Available).

`read.table(...)`

El archivo **03_datos_01.txt** contiene algunos datos sobre educación, recopilados para la Encuesta Nacional de Hogares (ENAHO) en el primer trimestre del 2024. La descripción de las variables puede consultarse en el diccionario.

1. Lea este archivo en RStudio asignándole el nombre datos01
2. Visualice el objeto leído directamente y usando la función View
3. ¿Qué estructura de datos es datos01? ¿Cuántas filas y columnas tiene?

`read.table(...)`

El archivo **03_datos_02.txt** contiene los mismos datos que el archivo anterior pero con una diferencia de forma.

1. Intente leer este archivo en RStudio usando el mismo código que la pregunta anterior
2. Lea este archivo correctamente y asígnele el nombre datos02
3. Seleccione los registros correspondientes a personas con educación superior universitaria completa

`read.table(...)`

El archivo **03_datos_03.txt** contiene los mismos datos que el archivo anterior pero con una diferencia de forma.

1. Lea este archivo correctamente.
2. Seleccione los registros correspondientes a las personas cuya lengua materna es el castellano y que tengan secundaria incompleta

`read.table(...)`

El archivo **03_datos_04.txt** contiene datos acerca de la lengua materna y nivel educativo alcanzado así como otros de ubicación, para los distritos de Ate, Chaclacayo y Lurigancho.

1. Lea solamente las 40 primeras filas de datos de este archivo, colocando los siguientes nombres de columnas: CONGLOME, VIVIENDA, HOGAR, PERSONA, UBIGEO, LENGUA, EDUCACION.
2. Seleccione aquellas personas que viven en Chaclacayo o que tengan secundaria completa.

read_tsv

Vuelva a leer los archivos anteriores usando la función read_tsv del paquete readr.

1. Instale y cargue el paquete readr
2. Almacene los data frames como datos01a, datos02a, datos03a y datos04a.
3. ¿Qué diferencias encuentra respecto a datos01, datos02, datos03 y datos04?

`read.csv(...)`

Argumento	Tipo	Descripción
file	Character Obligatorio	Archivo de datos a ser leído. Va entre comillas. Incluir su extensión
header	Logical Opcional	`TRUE` si la primera línea contiene los nombres de las variables. Por defecto se asume que es **`TRUE`.**
sep	Character Opcional	Caracter que indica el separador entre columnas. Va entre comillas, por defecto asume que es **`sep = ","`**.
dec	Character Opcional	Caracter que indica el separador de la parte entera y decimal en los números, por defecto asume **`dec = "."`**

`read.csv(...)`

Argumento	Tipo	Descripción
col.names	Vector Opcional	Vector que indica el nombre de las columnas
row.names	Vector Opcional	Vector que indica el nombre de las filas. También se puede indicar que una columna contiene estos nombres.
nrows	Integer Opcional	Número máximo de filas a leer
skip	Integer Opcional	Número de filas al inicio del archivo que no serán leídas
na.strings	Character Opcional	Caracter que indica cómo se declaran los valores perdidos (Not Available).

`read_csv(...)`

Argumento	Tipo	Descripción
file	Character Obligatorio	Archivo de datos a ser leído. Va entre comillas. Incluir su extensión
col_names	Logical Opcional	`TRUE` si la primera línea contiene los nombres de las variables. Por defecto se asume que es **`TRUE.`** También permite indicar el nombre de las columnas
n_max	Integer Opcional	Número máximo de filas a leer
skip	Integer Opcional	Número de filas al inicio del archivo que no serán leídas
na	Character Opcional	Caracter que indica cómo se declaran los valores perdidos (Not Available).

read.csv(...) y read_csv(...)

El archivo **03_datos_05.csv** contiene algunos datos recopilados por la ENAHO en Lima Metropolitana y Callao en el mes de marzo 2024.

1. Lea este archivo usando la función read.csv y almacénelo como datos05a.
2. Lea este archivo usando la función read.table y almacénelo como datos05b.
3. Use la función identical para verificar que datos05a y datos05b son iguales.
4. Lea este archivo usando la función read_csv del paquete readr y almacénelo como datos05c.
5. Compare los data frames leídos

`read.csv(...)` y `read_csv(...)`

El archivo **03_datos_06.csv** contiene todos los datos recopilados por la ENAHO para Lima Metropolitana y Callao en el mes de marzo 2024.

1. Lea el archivo de datos usando la función read.csv y almacénelo como datos06a.
2. Lea el archivo de datos usando la función read_csv y almacénelo como datos06b.
3. Filtre los datos de las personas que hablan quechua, aimara u otra lengua nativa y almacénelo en datos06c.

`read.csv2(...)`

Argumento	Tipo	Descripción
file	Character Obligatorio	Archivo de datos a ser leído. Va entre comillas. Incluir su extensión
header	Logical Opcional	`TRUE` si la primera línea contiene los nombres de las variables. Por defecto se asume que es **`TRUE`.**
sep	Character Opcional	Caracter que indica el separador entre columnas. Va entre comillas, por defecto asume que es **`sep = ";"`**.
dec	Character Opcional	Caracter que indica el separador de la parte entera y decimal en los números, por defecto asume **`dec = ","`**

`read.csv2(...)`

Argumento	Tipo	Descripción
col.names	Vector Opcional	Vector que indica el nombre de las columnas
row.names	Vector Opcional	Vector que indica el nombre de las filas. También se puede indicar que una columna contiene estos nombres.
nrows	Integer Opcional	Número máximo de filas a leer
skip	Integer Opcional	Número de filas al inicio del archivo que no serán leídas
na.strings	Character Opcional	Caracter que indica cómo se declaran los valores perdidos (Not Available).

`read.csv2(...)` y `read_csv2(...)`

El archivo **03_datos_07.csv** contiene una muestra aleatoria de 10000 datos del primer trimestre de la ENAHO 2024 - módulo 3.

1. Lea el archivo de datos usando la función read.csv2 y almacénelo como datos07a.
2. Lea el archivo de datos usando la función read.csv y almacénelo como datos07b.
3. Lea el archivo de datos usando la función read.table y almacénelo como datos07c.
4. Lea el archivo de datos usando la función read_csv2 y almacénelo como datos07d.
5. Lea el archivo de datos usando la función read_delim y almacénelo como datos07e.

No se nace sabiendo, se aprende practicando...

Ingrese a la web del INEI, seleccione ENAHO, 2024, Trimestre 1, y descargue el archivo CSV.

Lea el archivo usando las distintas funciones vistas en clase, y almacene los datos como datos08a, datos08b, etc.

Compare los data frames

Practique filtros simples, usando el operador “y”, y usando el operador “o”.

`read_xls(...)`, `read_xlsx(...)`, `read_excel(...)`

Argumento	Tipo	Descripción
path	Character Obligatorio	Archivo de datos de extensión .xls o .xlsx a ser leído. Va entre comillas. Incluir su extensión
sheet	Character o integer Opcional	Indica el nombre o número de hoja a ser leída.
range	Character Opcional	Indica el rango de datos a ser leído, por ejemplo A1:C4

`read_excel(...)`

Argumento	Tipo	Descripción
col_names	Logical o vector character Opcional	TRUE para indicar si la primera fila contiene los nombres de columnas (o FALSE en caso contrario), o un vector character indicando el nombre de las columnas
n_max	Integer Opcional	Número máximo de filas a leer
skip	Integer Opcional	Número de filas al inicio del archivo que no serán leídas
na	Character Opcional	Caracter que indica cómo se declaran los valores perdidos (Not Available).

read_xlsx(...) y read_xls(...)

El archivo **03_datos_09.xlsx** contiene algunos datos acerca de la ENAHO en el primer trimestre del 2024, módulo 3, en el distrito de San Juan de Lurigancho, así también el archivo **03 - DATOS - 09.xls**.

1. Instale y cargue el paquete readxl.
2. Lea el archivo 03 - DATOS - 09.xlsx usando la función read_excel almacénelo como datos09a.
3. Lea el archivo 03 - DATOS - 09.xlsx usando la función read_xlsx almacénelo como datos09b.
4. Lea el archivo 03 - DATOS - 09.xls usando la función read_xls almacénelo como datos09c.

`read_xlsx(...)` y `read_xls(...)`

El archivo **03_datos_10.xlsx** contiene los mismos datos de la pregunta anterior, con algunos cambios de forma.

1. Lea el archivo 03_datos_10.xlsx usando los argumentos skip y na, y almacénelo como datos10a
2. Lea el archivo 03_datos_10.xlsx usando los argumentos range y na, y almacénelo como datos10b.
3. ¿Los data frames leídos son idénticos?
4. Seleccione las personas que hablan quechua y usaron internet el mes pasado.

`read_xlsx(...)` y `read_xls(...)`

El archivo **03_datos_11.xlsx** contiene los datos de la ENAHO enero 2024 módulo 3, en 4 hojas: LIMA, Selva, sierra, Costa.

1. Lea la hoja LIMA del archivo 03_datos_11.xlsx de tres maneras distintas, almacenando las lecturas en datos11_lima_a, datos11_lima_b y datos11_lima_c.
2. Lea la hoja Selva del archivo 03_datos_11.xlsx de dos maneras distintas, almacenando las lecturas en datos11_selva_a y datos11_selva_b