Metodología Cuantitativa

Clase 2: Introducción a Tidyverse

Escuela de Trabajo Social - Pontificia Universidad Católica de Chile

Sebastián Rojas Vergara

01 de septiembre, 2021

Contenidos

1. Presentar el paquete `dplyr` de tidyverse..
2. Presentar el flujo de trabajo con proyectos de RStudio.
3. Practicar el procesamiento de datos con dplyr para la manipulación de filas y columnas.

I. Repaso de la clase anterior

Clase anterior

En la clase anterior revisamos:

- Cómo iniciar Rstudio y descripción de los 4 paneles (consola, enviroment, utilidades y script).
- Cómo comenzar un script en R y la instalación de paquetes.
- Concepto de **objeto** en R. Recomendaciones para nombrar objetos.
- Tipos y estructuras de datos con los comandos `class` y `typeof`.
- Presentación y construcción de vectores, matrices, data frames (tibbles) y listas.
- Cómo comentar, seccionar y guardar un script.

Tidyverse

- Es una colección de paquetes diseñadas para ciencia de datos. Todos los paquetes comparten una filosofía de diseño, gramática y estructura de datos. En particular, en esta clase veremos el paquete **`dplyr`** que está pensado en la manipulación de datos.

- También veremos el paquete **`haven`** que sirve para importar y exportar bases de datos en Stata y SPSS. Este ya viene instalado por **`tidyverse`**, pero debe cargarse con la función **`library`**.

Algunas funciones de `dplyr` I

Manipulación básica de columnas

- **`relocate()`**: reordena la posición de las variables en una base de datos.
- **`select()`**: selecciona y devuelve un conjunto de columnas.
- **`rename()`**: renombra columnas en una base de datos.

Manipulación básica de filas

- **`arrange()`**: reordena filas de un data frame.
- **`filter()`**: selecciona y devuelve un conjunto de filas según una o varias condiciones lógicas.
- **`slice()`**: selecciona filas basadas en su posición.

Algunas funciones de `dplyr` II

Herramientas básicas de edición de datos

- **`recode()`**: permite el reemplazo de valores numéricos; también de variables `character` y `factor` basados en sus nombres. En el caso de vectores lógicos.
- **`if_else()`**: evaluación de condiciones, y asignación de valores.
- **`na_if()`**: convierte valores a NA según una condición.
- **`mutate()`**: añade nuevas variables o transforma variables existentes.
- **`summarise()`**: calcula un resumen a partir de funciones de R (media, mínimo, máximo, etc.).
- **`group_by()`**: agrupa filas según una o más variables. Para remover la agrupación, se debe utilizar posteriormente el comando **`ungroup()`**.
- **`rowwise()`**: agrupa resultados por filas. Muy útil en conjunto con **`mutate()`**.

Algunas funciones de `dplyr` III

Otras funciones relevantes

- **`across()`**: es una función más avanzada que permite realizar operaciones circulares (loops) en menos líneas de código y de manera más eficiente en términos computacionales. Con esta función, por ejemplo, se podría calcular la media para múltiples variables simultáneamente. Se utiliza en conjunto con **`mutate()`** y **`summarise()`**.

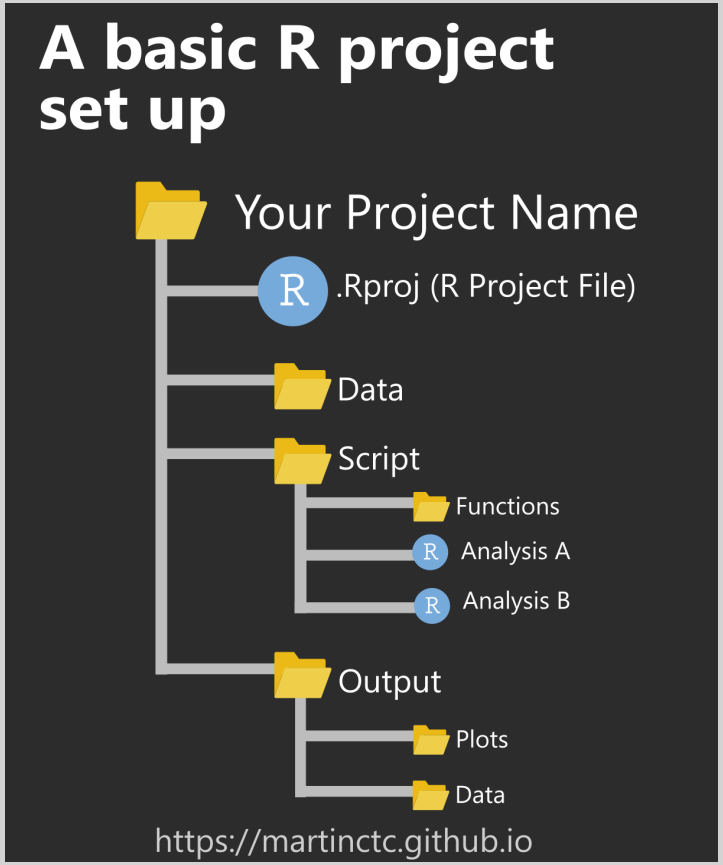
- **`case_when()`**: es una versión generalizada y más avanzada del comando **`if_else()`** y permite evaluar múltiples condiciones. Si los casos no calzan con ninguna de estas condiciones, R devolverá un valor NA (missing).

- **`count()`**: permite contar los valores para una o más variables. Se puede usar en conjunto con **`group_by()`**.

2. Trabajar con proyectos de RStudio

Directorios de trabajo y Proyectos

- El directorio de trabajo refiere a la ruta del computador donde R busca los archivos que le pedimos que lea y dónde guarda los archivos con los que trabajamos.
- Las rutas pueden causar problemas porque se escriben de manera diferente según el sistema operativo y porque es improbable que los usuarios tengan la misma ruta.
- Por esta razón, se trabajará con los archivos **Project** de Rstudio. Esto les aliviará muchísimo sufrimiento con R.
- Para crear un proyecto deben seleccionar ***File*** → ***New Project*** y escoger entre un nuevo directorio o uno ya existente. Usualmente será esta última opción, asumiendo que ya tienen la carpeta creada.



Un ejemplo de una configuración básica de un proyecto en RStudio.

3. Ejercicios con dplyr y condiciones lógicas

Ejercicio

Si tienes problemas con un comando o paquete, puedes escribir **`?nombre_comando`** o **`?nombre_paquete`** y se abrirá una pantalla de ayuda.

1. Carga y/o instala según corresponda los siguientes paquetes: tidyverse (dplyr) y haven.
2. Utilizando la función **`read_stata`** del paquete haven, importa los datos de ejemplo llamada *"02_Clase 2_Muestra_Casen2020"*.

A continuación realizaremos ejercicios con los siguientes comandos de dplyr.

- **`read_stata()`**
- **`glimpse()`**
- **`relocate()`**
- **`select()`**
- **`rename()`**
- **`count()`**
- **`group_by()`**
- **`filter()`**
- **`arrange()`**
- **`slice()`**

Condiciones lógicas

Para la programación en R se requiere manejar condiciones logícas. A continuación, se presentan las más relevantes.

Símbolo	Significado
==	Igualdad. No debe confundirse con =, que es un operador de asignación
!=	Distinto
!	Negación de una expresión lógica
>	Mayor que
>=	Mayor o igual que
<	Menor que
<=	Menor o igual que
&	Operador “y”. Todas las condiciones evaluadas deben ser verdaderas para que la expresión completa lo sea
|	Operador “o”. Al menos una condición debe ser verdaderas para que la expresión completa lo sea

¿Y ahora qué sigue?

Hay muchísimo material para revisar en R. Cada día se aprenden nuevas cosas y es absolutamente normal tener problemas o quedarse atascado con un código. Para nuestra suerte, la comunidad es muy activa. Algunas referencias:

- **Libro R para Ciencia de Datos**. Traducción en español del libro "R for Data Science" https://es.r4ds.hadley.nz/.
- **RStudio cheatsheets**. Una serie de "torpedos" con las funciones más útiles de cada paquete https://www.rstudio.com/resources/cheatsheets/.
- **R Ladies - Chile**. Es una organización que promueve la diversidad de género en la comunidad de R. Cuenta con grupos en Santiago, Valparaíso, Concepción y Talca. https://www.meetup.com/rladies-scl/.
- **Stackoverflow**. Página web que funciona en modalidad de foro donde otros usuarios responden tus dudas. Se usa para múltiples lenguajes de programación, incluido R. Muchísimas de las dudas que tengas ya tendrán respuesta aquí.
- Y muchas otras más: AnalizaR Datos Políticos; Manual de R Sociología Universidad de Chile - Estadística Descriptiva; R Bloggers y páginas de Twitter como @rfunctionaday.

¡Gracias!

Metodología Cuantitativa

Clase 2: Introducción a Tidyverse

Escuela de Trabajo Social - Pontificia Universidad Católica de Chile

Sebastián Rojas Vergara

01 de septiembre, 2021