OPSO79-1-UCSH2021

Investigación por encuestas; principales encuestas en Chile y como importarlas en R. Bloque práctico (5b)

10/09/2021

Introducción

Sesión anterior comenzamos a importar data a R desde archivos en nuestro pc

Para hacerlo fue central el uso de los `R Project`. Con estos creamos una carpeta o directorio en nuestro computador.

Desde esta ruta importamos y exportamos datos para cada proyecto, tarea o prueba que hacemos.

Si funciona bien. En carpeta que creamos aparecerá un archivo `.RProj` que fija la ruta y que debemos abrir cada vez que queramos trabajar en nuestro proyecto.

¡OJO! Este archivo es ligero, solo fija la ruta, no guarda la sesión de R ni lo que hayamos hecho.

Todo eso se guarda como código en archivo `.rmd` o `.r`

El formato para guardar datos en R es `.rds` y `.Rdata` (la data de Afganistán estaba en este formato), pero por ahora no usaremos esto.

Introducción

Hoy seguiremos usando los `.RProj` para importar data.

Veremos cuáles son las principales data frames que en sociología se suelen usar para responder a preguntas de investigación (sesgado):

- Encuesta ENE, ESI y ENCLA
- Encuesta CASEN
- Encuesta Discapacidad
- Latinobarómetro

Sobre todo, veremos como cargarlas y visualizarlas en R desde los distintos formas en los que se encuentran liberadas (excel, spss y stata).

La lógica es la misma para los distintos formatos.

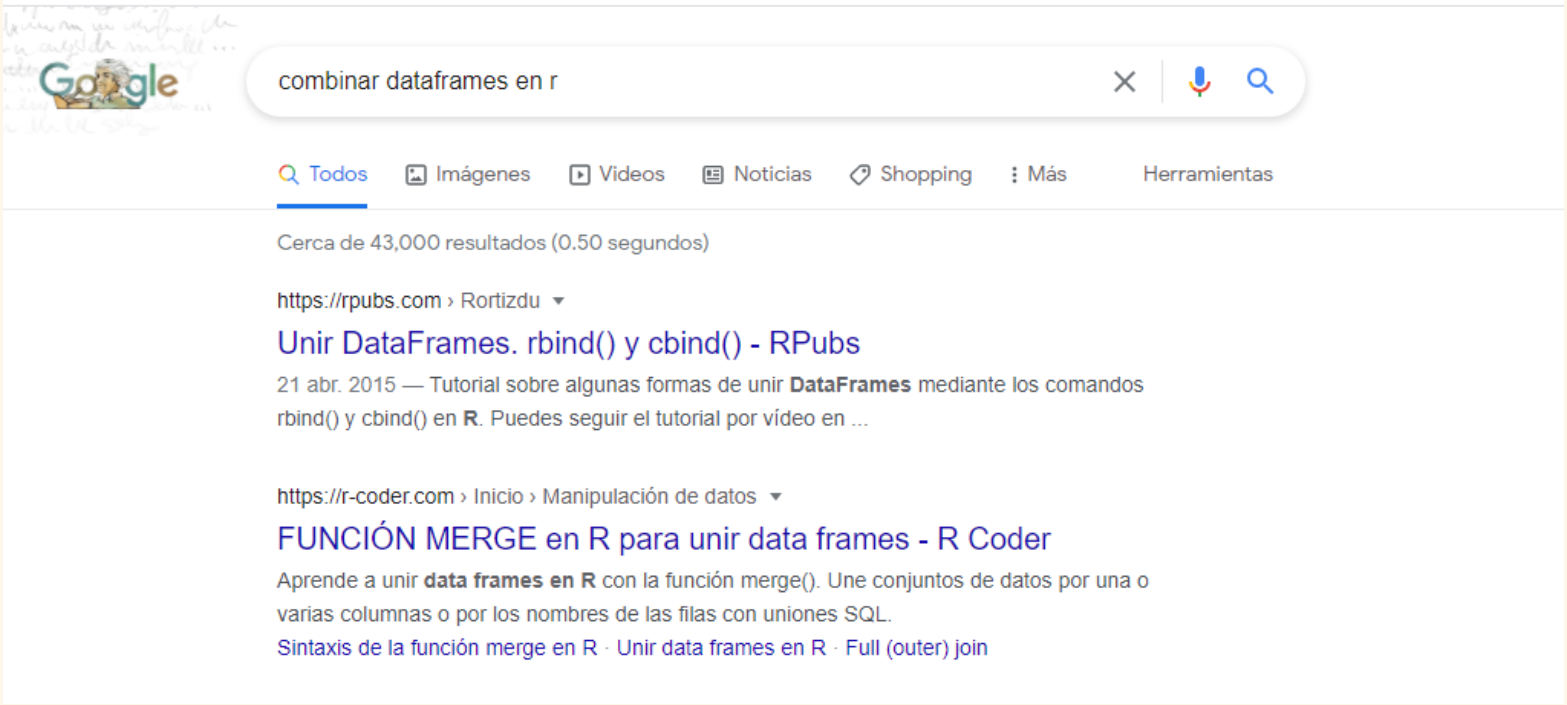
Antes veremos recomendaciones para "pedir ayuda" cuando no sabemos como hacer algo en R (como cargar datos en un formato específico)

Pedir ayuda en R

Pedir ayuda en R

Ningún libro, clase o tutorial te dará todas las herramientas para solucionar los **problemas** a los que te enfrentarás en R

Si en algún momento ya no puedes avanzar, empieza buscando en Google:



Pedir ayuda en R

A medida que el tema es más complejo, menos respuestas habrán en español.

En inglés, el 99% de las veces encontrarás respuestas para lo que buscas.

Muchas de las preguntas que hagas, otros ya las habrán hecho antes que tú.

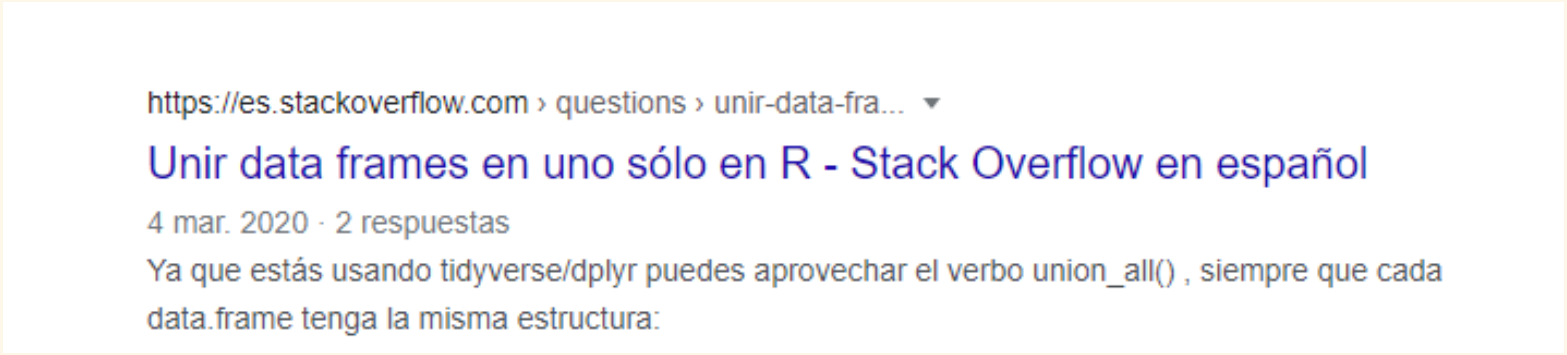
Sin embargo, eventualmente podría ser necesario preguntar...

Hay lugares y buenas prácticas para hacerlo.

Pedir ayuda en R

Desde *google* siempre nos saldrán páginas similares de foros y consultas en R.

La más confiable:



También muy recomendado *datanovia*, *r-bloggers*, las documentaciones de los paquetes, secciones de libros, etc.

Pedir ayuda en R

La clave está en saber preguntar en google...

Sin embargo, a veces podemos no saber que preguntar, ¿por qué nos aparece un **error** en el código?

```
> guaguas<-guaguas::guaguas
> select(guaguas,n,nombre)
Error in select(guaguas, n, nombre) :
  no se pudo encontrar la función "select"
> |
```

Detenerse a leer el error e intentar comprenderlo es buena práctica. Si no consigues comprender, **¡cópialo y pégalo en google!**

Muchos otros ya lo copiaron y pegaron. Por lo general alguien ya les respondió.

Si el error te aparece en español, con la siguiente función te aparecerá en inglés la próxima vez

```
Sys.setenv(LANGUAGE = "en")
```

Pedir ayuda en R

Aprender a solucionar las "panas" de manera **autónoma** es cuando uno/a más aprende R.

Sin embargo, estamos en un curso y como equipo decente esperamos ayudarlos y acompañarlos en sus aproximaciones a R.

Por tanto, si no consiguen resolver sus problemas consultando en *google* pueden preguntar vía mail (como hasta ahora lo han hecho)

La mejor manera de preguntar es con **ejemplos reproducibles** (*reprex*):

- explicitar de forma clara el problema (puede ser una foto del error)
- adjuntar la data sobre la que trabajas (o una parte de ella si es muy pesada)
- adjuntar código con el que estás trabajando (*.R* o *.rmd*)

Estos puntos son la base para preguntar en foros.

Principales encuestas en ciencias sociales y como importarlas en R

Encuesta ENE (INE)

La Encuesta Nacional de Empleo (ENE) se aplica desde inicios de 2010.

Es una encuesta a hogares que se aplica en viviendas particulares ocupadas.

Los informantes son los integrantes del hogar (o un representante)

La ENE clasifica y caracteriza a todas las personas en edad de trabajar (15 años y más).

El objetivo principal de la ENE es medir la **tasa de desocupación**.

Sin embargo, permite medir muchas otras cosas: principales ocupaciones por región, niveles de informalidad, diferencias por sexo, horas de trabajo, trabajo desde el hogar, etc.

La publicación de datos es mensual en la web (último día hábil del mes)

ENE: web

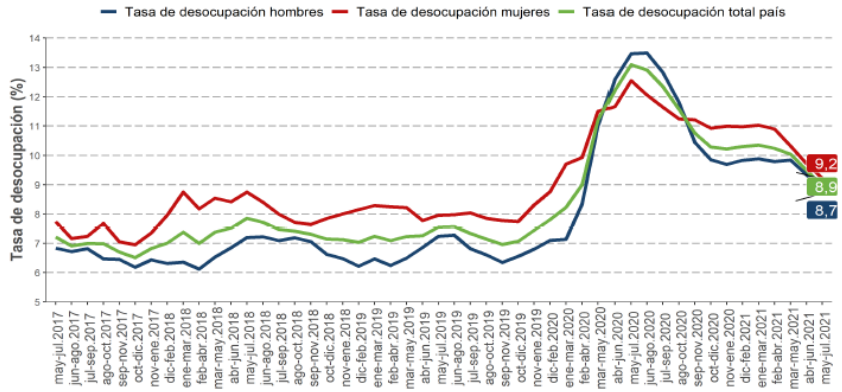

Ocupación y desocupación
Acá podrá encontrar los resultados y la documentación metodológica sobre la Encuesta Nacional de Empleo (ENE), además puede descargar las bases de datos y acceder a herramientas tales como el Banco de Datos ENE y Datos Abiertos del Mercado Laboral.
Para acceder a mayor información, además de los cuadros estadísticos, puede utilizar las siguientes herramientas: Banco de Datos ENE, Datos Abiertos .Stat o puede descargar directamente las bases de datos en formatos CSV, STATA o SPSS.
CUADROS ESTADÍSTICOS
BOLETINES
PUBLICACIONES Y ANUARIOS
DOCUMENTOS DE TRABAJO
METODOLOGÍAS
COMITÉS Y NOTAS TÉCNICAS
BASES DE DATOS
FORMULARIOS
METADATOS
Seleccione un elemento para navegar

ENE: boletín

- Los ocupados ausentes[2], que representaron el 9,5% del total de ocupados, decrecieron 42,0% (equivalente a 562.819 personas) en doce meses. Los trabajadores acogidos a la Ley de Protección al Empleo se encuentran en esta categoría.
- La población fuera de la fuerza de trabajo disminuyó 7,8% en doce meses, influida por la fuerza de trabajo potencial, que son personas que, debido a la contingencia sanitaria, en su mayoría no estaban buscando un trabajo, pero estaban disponibles para trabajar.
- La tasa de desocupación ajustada estacionalmente fue 8,7%, retrocediendo 0,7 pp. con respecto al trimestre móvil anterior.

Evolución tasa de desocupación, según sexo, total país

trimestres móviles



Fuerza de trabajo	8.948,33
Ocupados	8.148,95
Ocupados informales	2.191,97
Desocupados	799,38
Inactivos	6.975,51
Variaciones a 12 meses	
Fuerza de trabajo	9,9%
Ocupados	15,2%
Ocupados informales	38,9%
Desocupados	-25,0%
Inactivos	-7,8%
Tasas analíticas	
Tasa de desocupación con iniciadores disponibles (SU1)	9,3%
Tasa combinada de desocupación y tiempo parcial involuntario (SU2)	15,2%
Tasa combinada de desocupación y fuerza de trabajo potencial (SU3)	19,2%
Tasa de presión laboral	14,2%
Informalidad laboral	
Tasa de ocupación informal	26,9%
Tasa de ocupación en el sector informal	17,6%

ENE: tabulados

SERIE: OCUPADOS POR GRUPO DE OCUPACIÓN /1 Y SEGÚN TRIMESTRE /2

Nacional

Ambos sexos

Año	Trimestre	Ocupados (total)		Directores, gerentes y administradores		Profesionales, científicos e intelectuales		Técnicos y profesionales de nivel medio		Personal de apoyo administrativo		Trabaj los s vend com me
		nota	en miles	nota	en miles	nota	en miles	nota	en miles	nota	en miles	nota
2020	Ago - Oct		7.667,66		298,62		1.236,58		982,42		456,78	
2020	Sep - Nov		7.916,72		311,95		1.278,01		994,54		468,94	
2020	Oct - Dic		8.026,22		299,89		1.252,62		997,78		462,50	
2020	Nov - Ene		8.121,42		305,79		1.253,47		1.013,81		513,52	
2021	Dic - Feb		8.167,62		296,07		1.224,32		1.029,71		523,10	
2021	Ene - Mar		8.148,21		310,74		1.231,27		1.001,13		519,36	
2021	Feb - Abr		8.104,13		316,45		1.227,96		997,48		477,27	
2021	Mar - May		8.041,11		336,48		1.239,17		988,88		479,98	
2021	Abr - Jun		8.041,19		325,06		1.265,97		988,94		481,36	
2021	May -Jul		8.148,95		330,85		1.279,45		993,48		489,07	

Fuente: Encuesta Nacional de Empleo, INE-Chile.

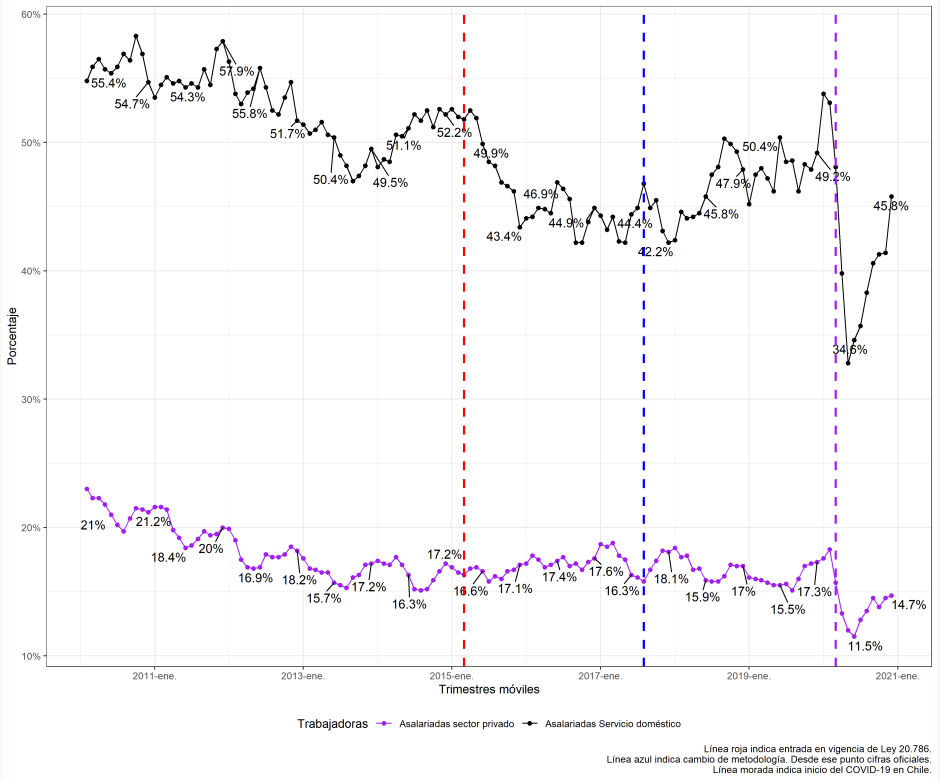
/1 Según Clasificador Chileno de Ocupaciones, CIUO 08.cl, adaptación de la Clasificación Internacional Uniforme de Ocupaciones (CIUO 08) c
https://www.ine.cl/docs/default-source/publicaciones/2018/ciuo-08.cl-clasificador-chileno-de-ocupaciones.pdf
/2 Serie disponible desde trimestre enero-marzo de 2017. Este tabulado reemplazó al de grupo de ocupación basado en CIUO-88, cuyo últimc
/3 A contar de la submuestra de enero 2020 el cuestionario central de la ENE permite la alternativa No sabe/No responde.

a: estimación poco fiable (coeficiente de variación mayor a 15% y menor o igual a 30%. En el caso de estimaciones de razón, si no cumple cor
b: estimación no fiable (número de casos muestrales menor a 60, grados de libertad menores a 9 o coeficiente de variación mayor a 30%)

ENE: microdatos

Todo esto son resúmenes agregados de la encuesta, estandarizados.

Tiene mucha potencia, pero también hay muchas cosas más que no se analizan:



ENE: micrdatos

Para estas y otras cosas recurrimos a los microdatos de la encuesta (pestaña **bases de datos**).

Datos en formato excel (.csv), stata (.dta) y spss (.sav).

Pasos a seguir:

- Crear R Project
- Descargar los datos del último trimestre ENE (formato spss)
- Guardarlos en carpeta del nuevo R proyect

Como puede tardar un poco, pueden descargar submuestra acá:

descarga muestra ene-2021-06-mjj.sav

Importar SPSS

- Paquete de R para importar y exportar archivos "externos" a R, como SPSS, Stata y SAS.
- Es una paquete "familiar" de `dplyr` (`tidyverse`), por lo que dialoga bien con otras funciones que utilizaremos y se encuentra permanentemente actualizado.
- Su creador es Hadley Wickham.


```
#install.packages("haven")
library(haven)
```

ENE: micrdatos

Cargar la data:

```
ene <- haven::read_spss("data/ene_de_muestra.sav")
```

Revisar lo cargado

```
ene[1:7,1:3]
```

```
## # A tibble: 7 x 3
##   ano_trimestre      mes_central                region
##           <dbl>        <dbl+lbl>              <dbl+lbl>
## 1          2021 6 [Mayo – Julio] 15 [Arica y Parinacota]
## 2          2021 6 [Mayo – Julio] 13 [Metropolitana]
## 3          2021 6 [Mayo – Julio]  9 [La Araucanía]
## 4          2021 6 [Mayo – Julio] 14 [Los Ríos]
## 5          2021 6 [Mayo – Julio] 11 [Aysén]
## 6          2021 6 [Mayo – Julio]  7 [Maule]
## 7          2021 6 [Mayo – Julio] 14 [Los Ríos]
```

ENE: micrdatos

Volvemos a ocupar paquete `sjmisc` para buscar variables

```
library(sjmisc)
find_var(ene,"horas")[1:10,3]
```

```
##  [1] "c2_1_3. Actividad principal: Total horas semanales trabajadas habitua
##  [2] "c2_2_3. Actividad secundaria: Total horas semanales trabajadas habitu
##  [3] "c3_3. Actividad principal: Total horas semanales contratadas o acorda
##  [4] "c4. ¿Le pagan habitualmente las horas extras en su actividad principa
##  [5] "c5. La semana pasada, ¿trabajó más horas que las habituales en su act
##  [6] "c6. ¿Cuántas horas más de las habituales trabajó la semana pasada?"
##  [7] "c7. La semana pasada, ¿trabajó menos horas que las habituales en su a
##  [8] "c8. ¿Cuántas horas menos de las habituales trabajó la semana pasada?"
##  [9] "c9. ¿Por qué razón trabajó un número de horas diferente a lo habitual
## [10] "c9_otro. Especifica c9 = Otras razones de un número de horas diferent
```

ENE: micrdatos

```
summary(ene$c2_1_3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    3.00   36.00   45.00   45.24   45.00  888.00     611
```

¿Existen diferencias por sexo?

```
frq(ene$sexo)
```

```
##
## Sexo (x) <numeric>
## # total N=1000  valid N=1000  mean=1.54  sd=0.50
##
## Value |  Label |   N | Raw % | Valid % | Cum. %
## ---------------------------------------------
##     1 | Hombre | 455 | 45.50 |   45.50 |  45.50
##     2 |  Mujer | 545 | 54.50 |   54.50 | 100.00
##  <NA> |   <NA> |   0 |  0.00 |    <NA> |   <NA>
```

ENE: micrdatos

Luego veamos el `summary()` para cada sexo

```
summary(ene[ene$sexo==1,]$c2_1_3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    4.00   40.00   45.00   45.53   45.00  888.00     235
```

```
summary(ene[ene$sexo==2,]$c2_1_3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    3.00   32.00   45.00   44.85   45.00  888.00     376
```

¿Por qué tanta diferencia en valor válido (NA's)?

```
table(ene$sexo,ene$categoria_ocupacion)
```

```
##
##       0   1   2   3   4   5   6   7
##   1 235   7  55 129  25   1   0   3
##   2 376   7  33  77  35  11   4   2
```

ENE: micrdatos

```
flat_table(ene,categoria_ocupacion,sexo)
```

```
##                                                    sexo Hombre Mujer
## categoria_ocupacion
## No corresponde                                              235   376
## Empleador                                                     7     7
## Cuenta propia                                                55    33
## Asalariado sector privado                                   129    77
## Asalariado sector público                                    25    35
## Personal de servicio doméstico puertas afuera                 1    11
## Personal de servicio doméstico puertas adentro                0     4
## Familiar o personal no remunerado                             3     2
```

ENE: micrdatos

```
flat_table(ene,categoria_ocupacion,sexo,margin = "col")
```

```
##                                                    sexo Hombre Mujer
## categoria_ocupacion
## No corresponde                                      51.65 68.99
## Empleador                                            1.54  1.28
## Cuenta propia                                       12.09  6.06
## Asalariado sector privado                           28.35 14.13
## Asalariado sector público                            5.49  6.42
## Personal de servicio doméstico puertas afuera        0.22  2.02
## Personal de servicio doméstico puertas adentro       0.00  0.73
## Familiar o personal no remunerado                    0.66  0.37
```

Importar excel (.csv)

Sí la base está en `spss` es mejor cargarla en ese formato.

La ventaja es que la data tiene etiquetas asociadas, lo que evita buscar valores en el **libro de códigos** de la encuesta

La desventaja es que la data pesa más.

Si queremos cargar las últimas 36 bases de datos de la ENE para ver la tendencia de las horas de trabajo, será más demoroso cargando `.sav` que `.csv`

La alternativa es descargar los archivos `.csv` y cargarlos

Importar excel (.csv)

Para el caso de `.csv` no es necesario cargar paquetes.

```
ene <- read.csv("data/ene_de_muestra.csv")
dim(ene)
```

```
## [1] 1000    1
```

¿Que pasó?

```
ene <- read.csv2("data/ene_de_muestra.csv")
dim(ene)
```

```
## [1] 1000  180
```

Los datos estaban separados por ";", no por ","

Importar excel (.csv)

Al ser más livianos también los podemos cargar directamente desde la web

(esto aplica a todos los formatos)

```
ene <- read.csv2("https://www.ine.cl/docs/default-source/ocupacion-y-des
```

Demora más que cargar lo descargado, por lo que no ejecutaré el código.

Pueden probar la línea de código con buena conexión a internet.

Importar .xlsx (excel)

Formato que no se suele usar para microdatos

Pero sí se usa mucho para cuadros o tabulados (agregados)

El archivo .xlsx puede tener más de una pestaña, de ahí la necesidad de una función específica

`readxl` también es de Wickham y del universo `tidyverse`

```
#install.packages(readxl)
library(readxl)
```

Importar .xlsx (excel)

Descargar cuadro de la ENE: "población-en-edad-de-trabajar-por-situación-en-la-fuerza-de-trabajo" (el primero)

La serie es como si hubiesen analizado y combinado cada microdato desde el 2010 a la fecha.

El archivo excel tiene muchas pestañas, por lo que hay que agregar más argumentos a la función genérica:

```
data <- función("direccion")
```

```
pob <- read_excel("data/población-en-edad-de-trabajar.xlsx",sheet = 2)
head(pob)[1:4]
```

```
## # A tibble: 6 x 4
##   `SERIE: POBLACIÓN EN EDAD DE TRABAJAR POR S~ ...2    ...3                .
##   <chr>                                        <chr>   <chr>               <
## 1 Nacional                                     <NA>    <NA>                <
## 2 Ambos sexos                                  <NA>    <NA>                <
## 3 <NA>                                         <NA>    <NA>                <
```

Importar .xlsx (excel)

El segundo problema es que hay celdas vacías al inicio.

Usar skip:

```
pob <- read_excel("data/población-en-edad-de-trabajar.xlsx",sheet = 2,
                  skip = 6)
head(pob)[1:6]
```

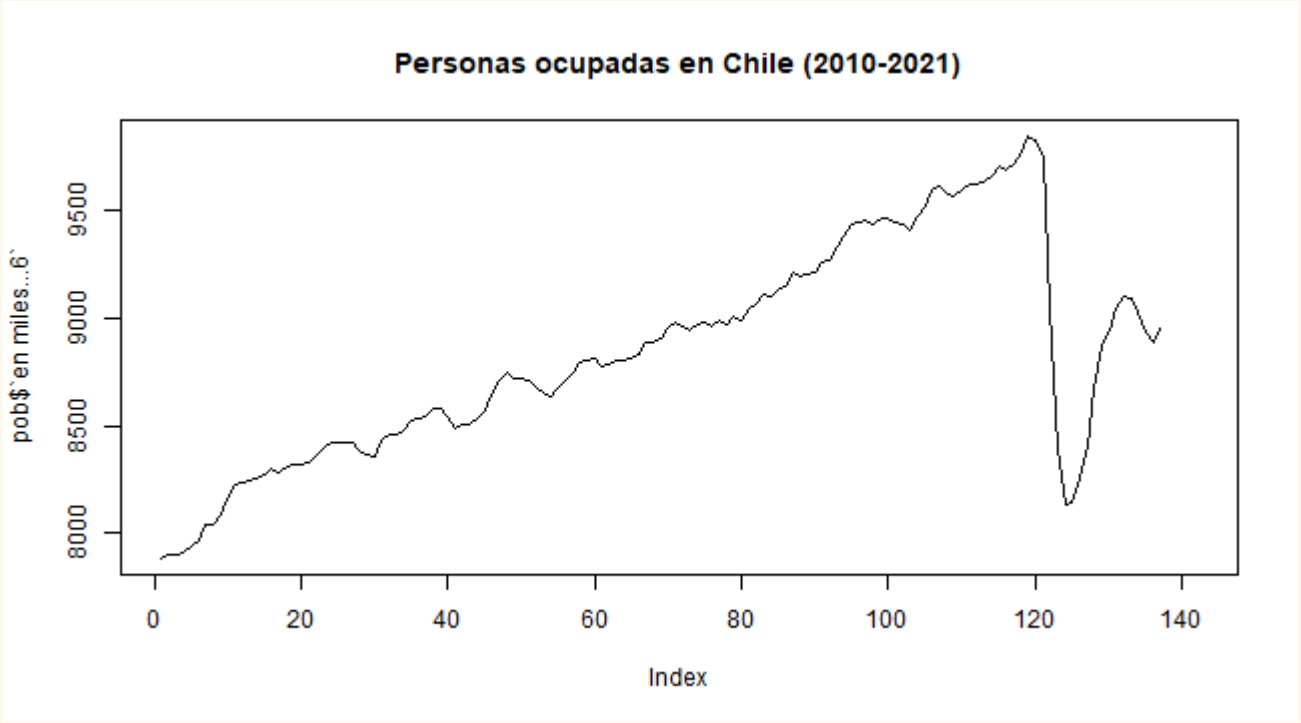
```
## # A tibble: 6 x 6
##   ...1  ...2      nota...3 `en miles...4` nota...5 `en miles...6`
##   <chr> <chr>     <lgl>             <dbl> <lgl>             <dbl>
## 1 2010  Ene - Mar NA               13218. NA                7884.
## 2 2010  Feb - Abr NA               13236. NA                7897.
## 3 2010  Mar - May NA               13253. NA                7900.
## 4 2010  Abr - Jun NA               13270. NA                7906.
## 5 2010  May -Jul  NA               13287. NA                7932.
## 6 2010  Jun - Ago NA               13305. NA                7961.
```

Hay claros problemas con los nombres y con variables vacías.

Hay funciones específicas para eso que veremos más adelante (`janitor`)

Importar .xlsx (excel)

```
plot(pob$`en miles...6`,type = "l",main="Personas ocupadas en Chile (201
```



Encuesta Laboral

La encuesta busca diagnosticar el estado y evolución de las condiciones de empleo y trabajo, de las relaciones laborales y de la igualdad de género en las empresas regidas por el Código del Trabajo en Chile.

- Su población objetivo corresponde a las empresas formales vigentes con cinco o más trabajadores contratados directamente.
- Cada empresa cuenta con tres unidades de observación (informante). Cada unidad constituye una base de datos: Empleadores, Autoaplicado, Trabajadores o Sindicatos.
- El año 2019 se aplicó la última versión: 3.670 empresas, 4 bases de datos.

Solo descarguemos la data de "sindicatos" (esta vez en formato `.dta` o stata).

Encuesta Laboral (Encla)

La lógica será la misma que en spss.

Descargar data y guardar en carpeta donde esté el R Project con el que estamos trabajando.

```
encla <- read_dta("data/bbdd-sindicatos-bp.dta")
```

Veamos la data

```
encla[1:5,1:3]
```

```
## # A tibble: 5 x 3
##     id_bp        a1        a2
##     <dbl> <dbl+lbl> <dbl+lbl>
## 1 4554856    2 [No]   NA
## 2 7550575    1 [Sí]    2 [No]
## 3 3363621    2 [No]   NA
## 4 4979613    1 [Sí]    1 [Sí]
## 5 2327403    1 [Sí]    1 [Sí]
```

Encuesta Laboral (Encla)

```
summary(encla$g2_3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0    50.0   110.5   278.4   241.2 20000.0
```

```
descr(encla$g2_3)
```

```
##
## ## Basic descriptive statistics
##
##  var    type
##   dd numeric
##
##  g2_3. ¿Cuántos hombres y mujeres se encuentran afiliados a su sindicato el
##     n NA.prc   mean     sd    se    md trimmed           range    iqr  skew
##  1172      0 278.45 851.36 24.87 110.5  278.45 19999 (1-20000) 191.25 14.65
```

¿Socios/as promedio por sector económico?

Encuesta Laboral (Encla)

Adelanto de lo que se viene:

```
library(dplyr)
descr(group_by(encla,agrupacion_actividad),g2_3)
```

Que sería como hacer esto para cada valor de `agrupacion_actividad` (13 valores)

```
# frq(encla$agrupacion_actividad)
summary(encla[encla$agrupacion_actividad==1,]$g2_3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   25.00   49.00   94.06  120.00  350.00
```

```
summary(encla[encla$agrupacion_actividad==2,]$g2_3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    34.0   127.0   208.0   317.6   424.0  1480.0
```

Para cerrar

Estas son solo algunas de las **principales encuestas** aplicadas en Chile. Hay muchísimas más que pueden ver en la página de estas y otras instituciones.

Pendiente Encuesta CASEN, Latinobarómetro y Encuesta Discapacidad (SENADIS).

Además, otra data no se produce por encuesta, sino que mediante el **registro de prensa** (Observatorio de Huelgas y Conflictos Sociales), los **registros administrativos** o mediante la **integración** de distintos datos (Banco Mundial, de donde venía la data Afganistán).

Sin embargo, toda la data que podría servir para la sociología y que se encuentra **liberada** para nosotros es solo una ínfima parte de toda la data que se produce en Chile.

¿Qué pasa con esta data que la producimos todos (aleatoriamente en el caso de las encuestas) pero que solo algunos pueden analizarla y sacarle provecho?

Este es uno de los pilares de la **ciencia abierta...** (tema post prueba 1)

Para practicar

- Crea un **R Project**, donde guardarás todos los archivos
- Crea un **RMarkdown** dentro del R Project para hacer la tarea
- Importa a R el microdato de la ENE del trimestre **OND 2019 (noviembre)**
- Genera un cuadro de resumen de `b14_rev4cl_canes`
- Cruza las variables `b14_rev4cl_canes` y `b1` (haz una tabla de contingencia).
- Crea la variable `pet` que tome valor 1 si `edad` es mayor o igual a 15, y 0 en otro caso.
- Crea la variable `ocu` (ocupados) que tome el valor 1 si la variable `cae_especifico` se encuentra en el rango (extremos incluidos) entre 1 y 7, y que tome el valor 0 en cualquier otro caso.
- ¿Cuantas personas `pet` y `ocu` hay en la muestra?
- ¿Cuál es la tasa de ocupación en noviembre de 2019? (`ocu / pet * 100`)
- Envía por correo el archivo `.html` resultante.

Recursos web utilizados

Xaringan: Presentation Ninja, de Yihui Xie. Para generar esta presentación.

Ilustraciones de Allison Horst

Para reforzar y seguir aprendiendo

Wickham, H. (2021) Pedir ayuda y aprender más

Bibliografía utilizada

Wickham, H. (2021). *R Para Ciencia de Datos*.