HARMONIZE

data4health

Harmonize training course - Nov 2025

This work was supported by the Wellcome Trust grant number 224694/Z/21/Z.

Agenda

Quick introduction - 20 min

Follow along - 15 min

Hands on - 2 hours

Feedback - 20 min

About the developers

Daniela Lührsen
Data Scientist

Raquel Martins Lana
Recognized Researcher

Carles Milà
Data Scientist

Emma Roberts
Data Scientist

Rachel Lowe
Principal Investigator

About you

Who has worked with health data before?

Who has worked with **messy** health data before?

What tools do you normally use to process health data?
e.g. software, packages, programing languages

Working with health data

Working with health data

My data has more than 50 columns, I must have so much information!

BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT
Muest	Resultado fin	CondiciÃ	Gravedad	Folio Ne	Semana defunciÃ³	Mes defunciÃ	AÃ±o defunciÃ	Fecha defunciÃ³	Edad fecha defunciÃ³	CIE 10 bÃ¡si	CIE 10	CIE 10	CIE 10	CIE 10	Diag. Final	Clasf. fin	Fuente detecciÃ³n	Confirmado por
No		Vivo	Sin signos de alarma														NotificaciÃ³n pasiva	
Si	Probable	Vivo													DENGUE	Probable	NotificaciÃ³n pasiva	CIÃnica
Si	Probable	Vivo															NotificaciÃ³n pasiva	
Si	Probable	Vivo															NotificaciÃ³n pasiva	
Si	Probable	Vivo															NotificaciÃ³n pasiva	
Si	Probable	Vivo															NotificaciÃ³n pasiva	
No		Vivo	Grave														BÃºsqueda institucional	
No		Vivo													ENFERMEDAD AGUDA GRAVE DE ETIOLOGÃ?A DESCONOCID	Confirmado	BÃºsqueda institucional	Laboratorio
No		Vivo													OTRA ENFERMEDAD	Confirmado	BÃºsqueda institucional	Laboratorio
No		Vivo													OTRA ENFERMEDAD	Descartado	BÃºsqueda institucional	
No		Vivo													OTRA ENFERMEDAD	Descartado	BÃºsqueda institucional	
Si	Probable	Vivo													OTRA ENFERMEDAD	Descartado	BÃºsqueda institucional	
Si	Probable	Vivo															BÃºsqueda institucional	
Si	Confirmado	Vivo															BÃºsqueda institucional	
Si	Probable	Vivo															BÃºsqueda institucional	
Si	Probable	Vivo															NotificaciÃ³n pasiva	
No		Vivo	Sin signos de alarma														BÃºsqueda institucional	
No		Vivo	Sin signos de alarma														BÃºsqueda institucional	
No		Vivo	Sin signos de alarma														BÃºsqueda institucional	
No		Vivo															BÃºsqueda institucional	
No		Vivo															BÃºsqueda institucional	
No		Vivo															BÃºsqueda institucional	
No		Vivo	Sin signos de alarma														BÃºsqueda institucional	
Si	Confirmado	Vivo															BÃºsqueda institucional	
Si	Probable	Vivo															BÃºsqueda institucional	
Si	Confirmado	Vivo															BÃºsqueda institucional	
Si	Probable	Vivo															BÃºsqueda institucional	
No		Vivo	Sin signos de alarma														BÃºsqueda institucional	
No		Vivo															NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
Si	Probable	Vivo															NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
Si	Probable	Vivo															NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
Si	Probable	Muerto		2017-105570	22	5	2017	31/05/2017	17 aÃ±os						OTRA ENFERMEDAD	Descartado	NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
Si	Probable	Vivo															NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
Si	Probable	Vivo															NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
Si	Confirmado	Vivo															NotificaciÃ³n pasiva	
Si	Probable	Vivo															NotificaciÃ³n pasiva	
Si	Probable	Vivo															NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
Si	Probable	Vivo	Grave														NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
Si	Probable	Vivo	Grave														NotificaciÃ³n pasiva	
Si	Probable	Vivo															NotificaciÃ³n pasiva	
Si	Probable	Vivo															NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
No		Vivo	Grave														NotificaciÃ³n pasiva	
No		Vivo	Sin signos de alarma														NotificaciÃ³n pasiva	
No		Vivo	Sin signos de alarma														NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
No		Vivo	Sin signos de alarma														NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
No		Vivo	Con signos de alarma														NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	
No		Vivo															NotificaciÃ³n pasiva	

Working with health data

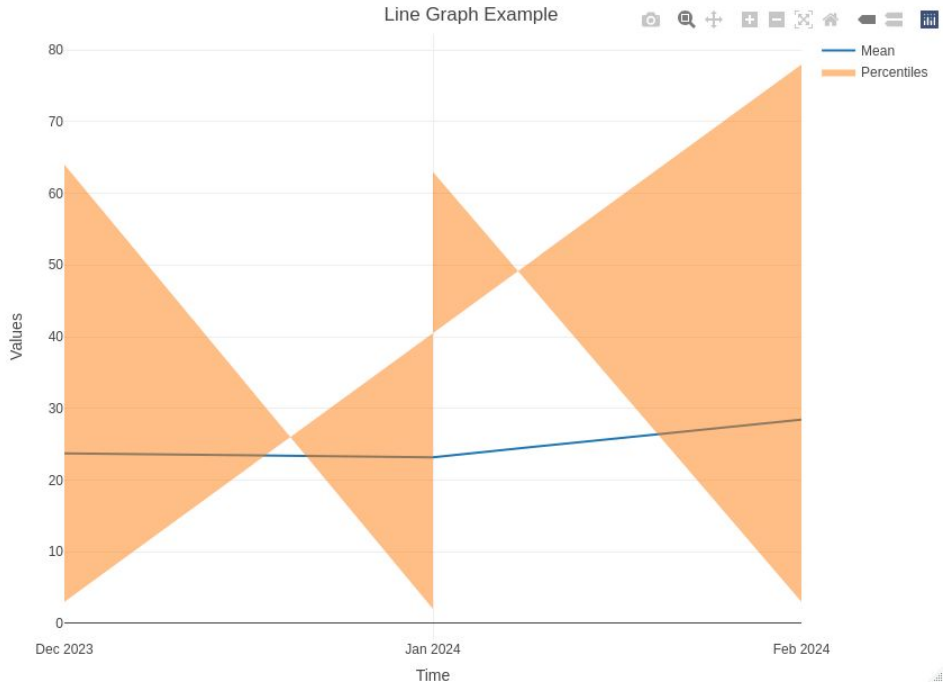
Working with health data

What the content of the column “age” looks like…

18	1.5 years
NA	19 DAYS
14mths	0.16700000000000001
5 dys	8.3000000000000004E-2
2 wks	6 Days
14 days	5 Mths
11 months	1 Wks
5.5/12	4 MTHS
6/52	3d
0.04	8m
0.666666666666667mnths	8 MONTHS
0.019 mnths	5M
10 WKS	9y

Working with health data

Working with health data



Pain points

Cleaning the data

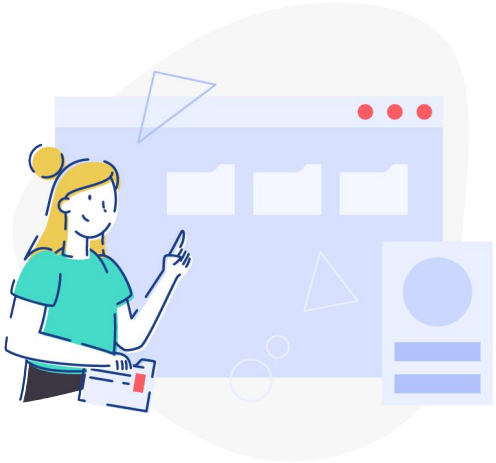
Repetitive data processing

Converting to/from epiweek

Visualising 

data4health - Workflow

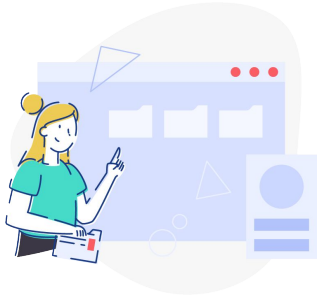
How to use data4health



Non-code experienced

Code experienced

Load



Upload your data here

Accepted file formats are .csv, .dbf, and .dbc.

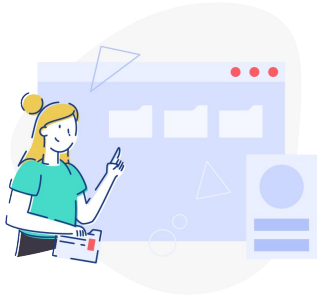
File should be in linelist format, each row represents a disease case and each column is a variable describing the case.

Browse... | No file selected | Load Test data

```
df <- d4h_load("data/input/example_data.dbc")

df <- d4h_load(c("data/Datos_2013_210.xlsx",
                 "data/Datos_2014_210.xlsx",
                 "data/Datos_2015_210.xlsx",
                 "data/Datos_2016_210.xlsx",
                 "data/Datos_2017_210.xls",
                 "data/Datos_2018_210.xls"))
```

Clean



Select a column

CS_SEXO

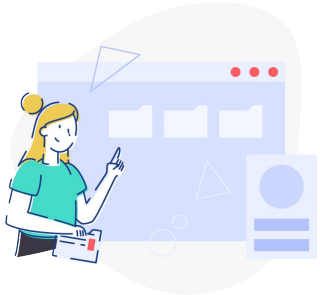
Current Category	New Categogry name
M	male
F	female
I	
NA	

Click here to show the code

Rename categories

```
df <- d4h_clean(data = df,
                cols_to_include = c("time", "region", "gender"),
                threshold_remove = 80,
                remove_rows_missing = "time",
                rename_columns =  c(time = "date",
                                    region = "municipality"),
                rename_categories = list(gender = c("mujer" = "female",
                                                    "M" = "male",
                                                    "F" = "female")),
                date_to_weekdate = "time",
                date_to_monthnumber = "time",
                date_to_yearnumber = "time")
```

Filter



Filter data if necessary

Column	Type2	Condition	Select Date Range
DT_NASC	Date	between	2024-07-01 to 2025-10-31

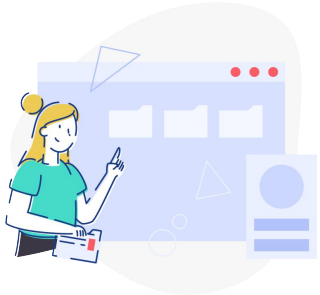
Column	Type2	Condition	Select category
CS_SEXO	character	include	M F

Click here to show the code

+ Add Filter | Apply Filter

```
df <- d4h_filter(data = df,
                 muni_code = list(include = c("region1", "region2")),
                 age = list(over = 18),
                 date = list(during = c("2018-01-01","2018-12-31")))
```

Aggregate



Select columns to aggregate

Temporal variable to aggregate: DT_SIN_PRI

Spatial variable to aggregate: ID_UNIDADE

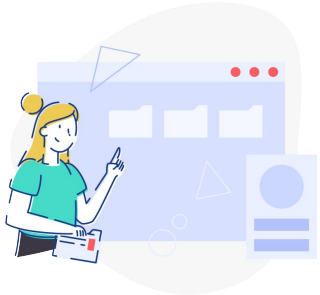
Optional additional variabbles to aggregate: CS_SEXO

Click here to show the code

Aggregate data

```
df <- d4h_aggregate(data = df,
                    space_col = c("muncipality"),
                    time_col = "date_month",
                    all_times = seq(from = min(date_month),
                                    to = max(date_month),
                                    by = "month"))
```

Save



Download Cleaned Data

```
d4h_save(data = df, name = "my_results")

d4h_save(data = df, name = "my_results2", extension = "rds")
```

Follow along - User Interface

1. Open Docker
2. Open 01_data4health_ui.Rmd

Follow along - Functions

- Open Docker
- Open 02_data4health_example.Rmd

Before we start with the hands-on

1. Try to solve it yourself
2. Ask for help
3. Check the solutions

What to look out for

Bugs that you encountered

New functionalities you would like to see added

Unclear error messages

How to give feedback - unclear error messages

Good error messages:

- Tell you clearly what is wrong
- Tell you how to correct

```
"Nothing to aggregate by. Please
provide a 'time_col', 'space_col'
or 'add_col'."
```

```
"'remove_rows_missing' must be NULL
or a single string (column name)."
```

```
"Could not find column
'notificaion_unit' in dataset."
```

Bad error messages:

- Unclear what went wrong
- Often uses unknown terms

```
Error: only defined on a data frame
with all numeric-alike variables
```

```
Error in df["DT SIN PRI"] : object
of type 'closure' is not
subsettable
```

Now it's your turn!

How can data4health improve?

Today you have used the very first version of data4health.

To keep improving and making the tool more useful to a bigger audience, we require the feedback from users.

Please fill in this form as detailed as possible.

https://tinyurl.com/HARMONIZE-feedback

Thank you!

Harmonize training course - Nov 2025

This work was supported by the Wellcome Trust grant number 224694/Z/21/Z.