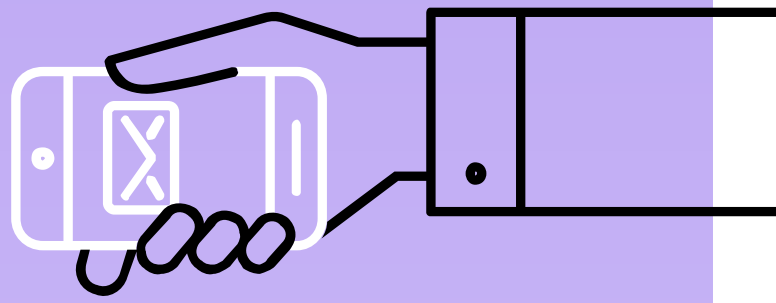
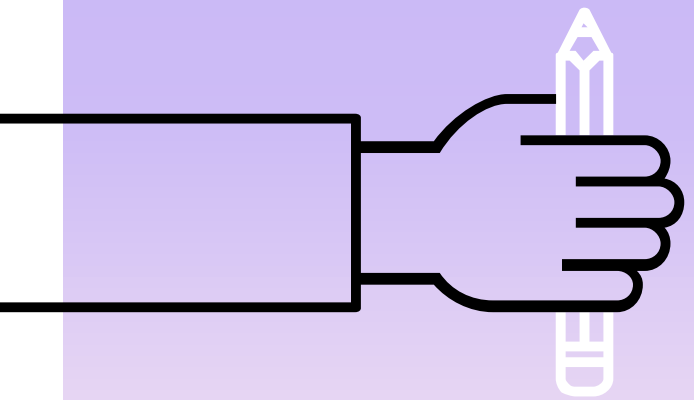
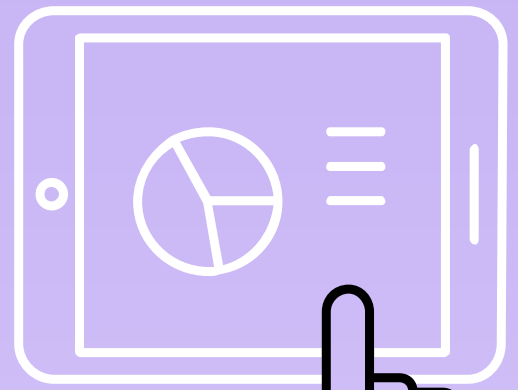
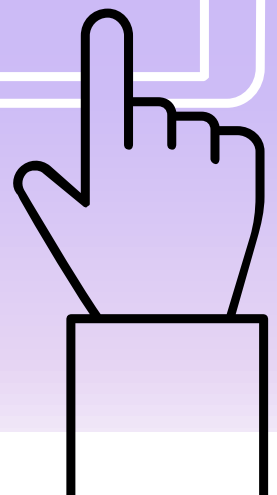


Data Science Academy Mexico

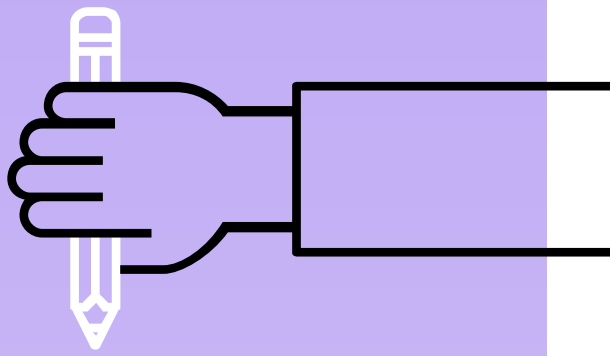
Limpieza y manipulación de datos en R





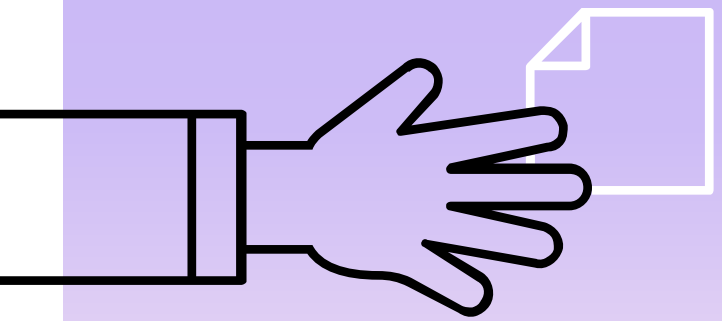


Febrero 21, 2019



Hola!

Presentación



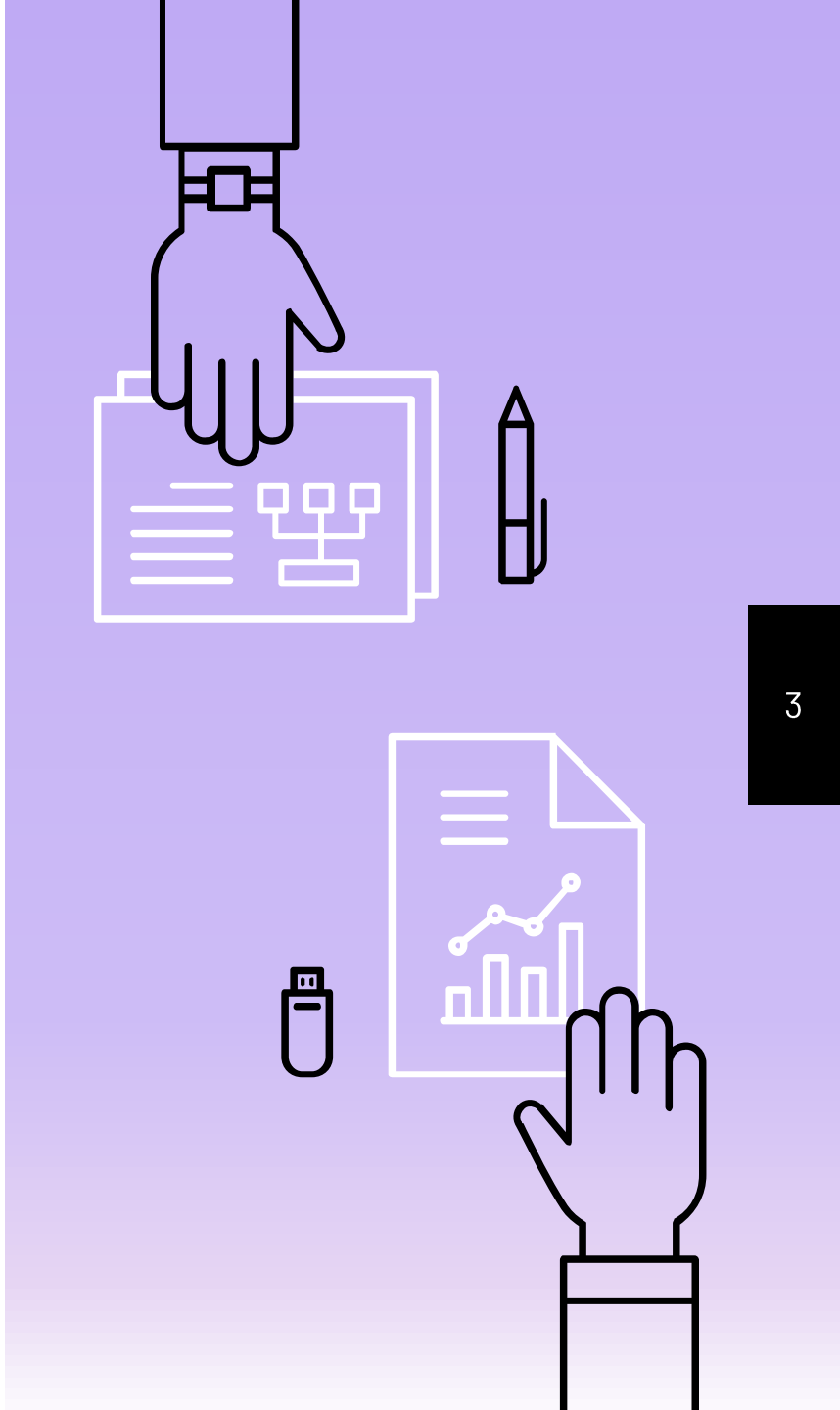
Rocío Maribel Ávila Ayala

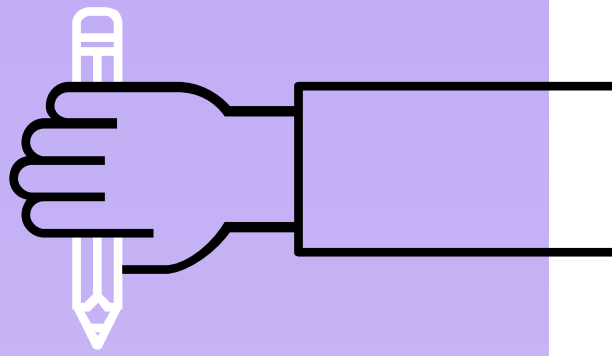
Formación Académica

- Actuaría (UNAM FES-Acatlán)
- Maestría en Estadística (CIMAT)

Experiencia profesional

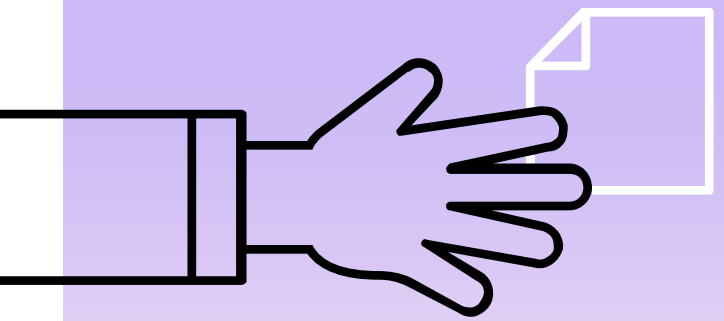
- Seguros AFIRME
- Las Quince Letras Solutions
- General Motors de México (Actual)





Introducción

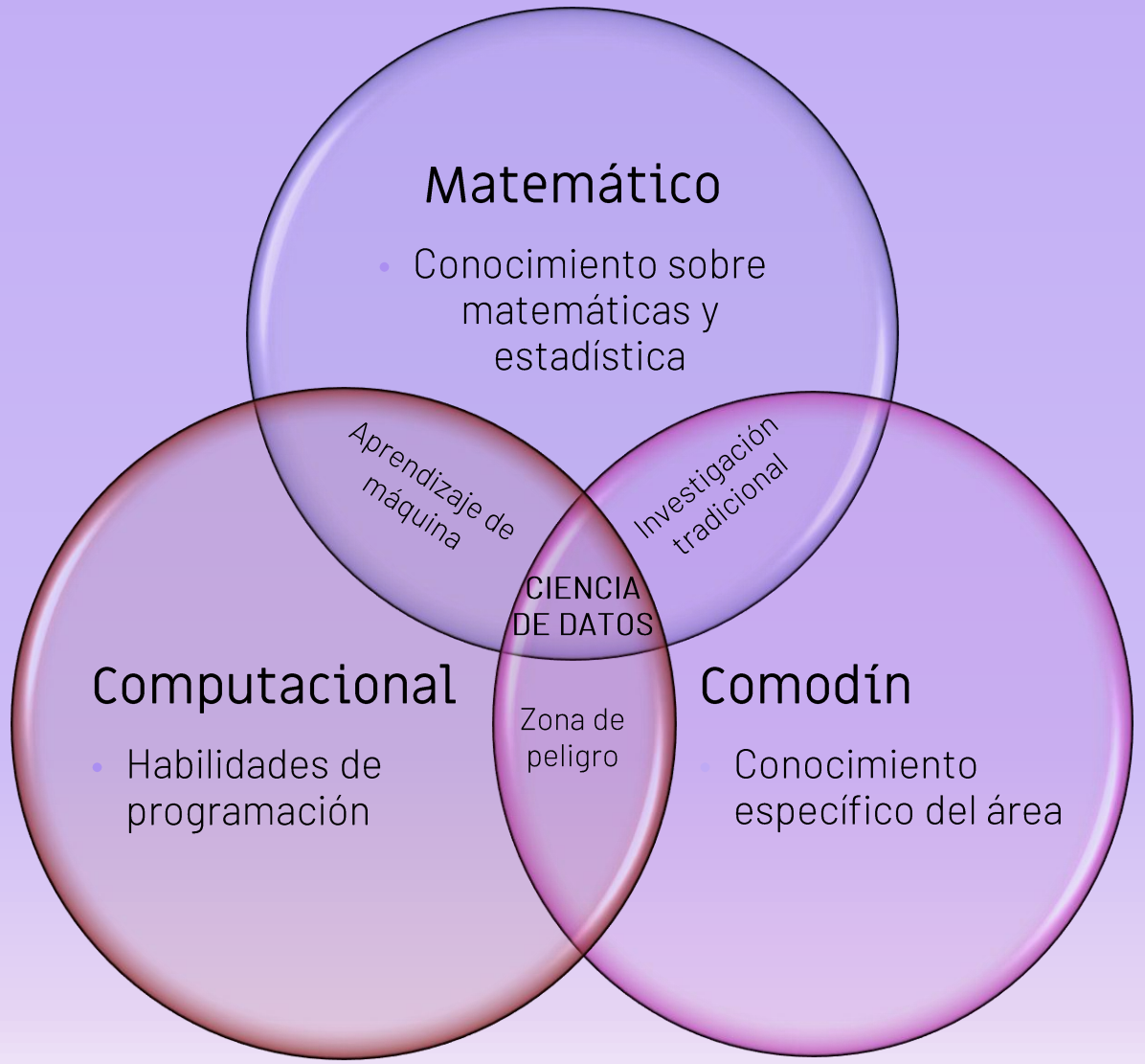
Importancia de la limpieza y manipulación de datos



CIENCIA de datos

Disciplina que nos permite transformar **datos en bruto** en **información**, con el objetivo de convertir esta última en **conocimiento** sobre un tema específico de interés

Pilares que fundamentan la Ciencia de Datos



CIENCIA de datos

Disciplina que nos permite transformar **datos en bruto** en **información**, con el objetivo de convertir esta última en **conocimiento** sobre un tema específico de interés

Datos en bruto

- Hechos desorganizados
- No proveen más información de lo que se ve a simple vista
- Múltiples fuentes

Información

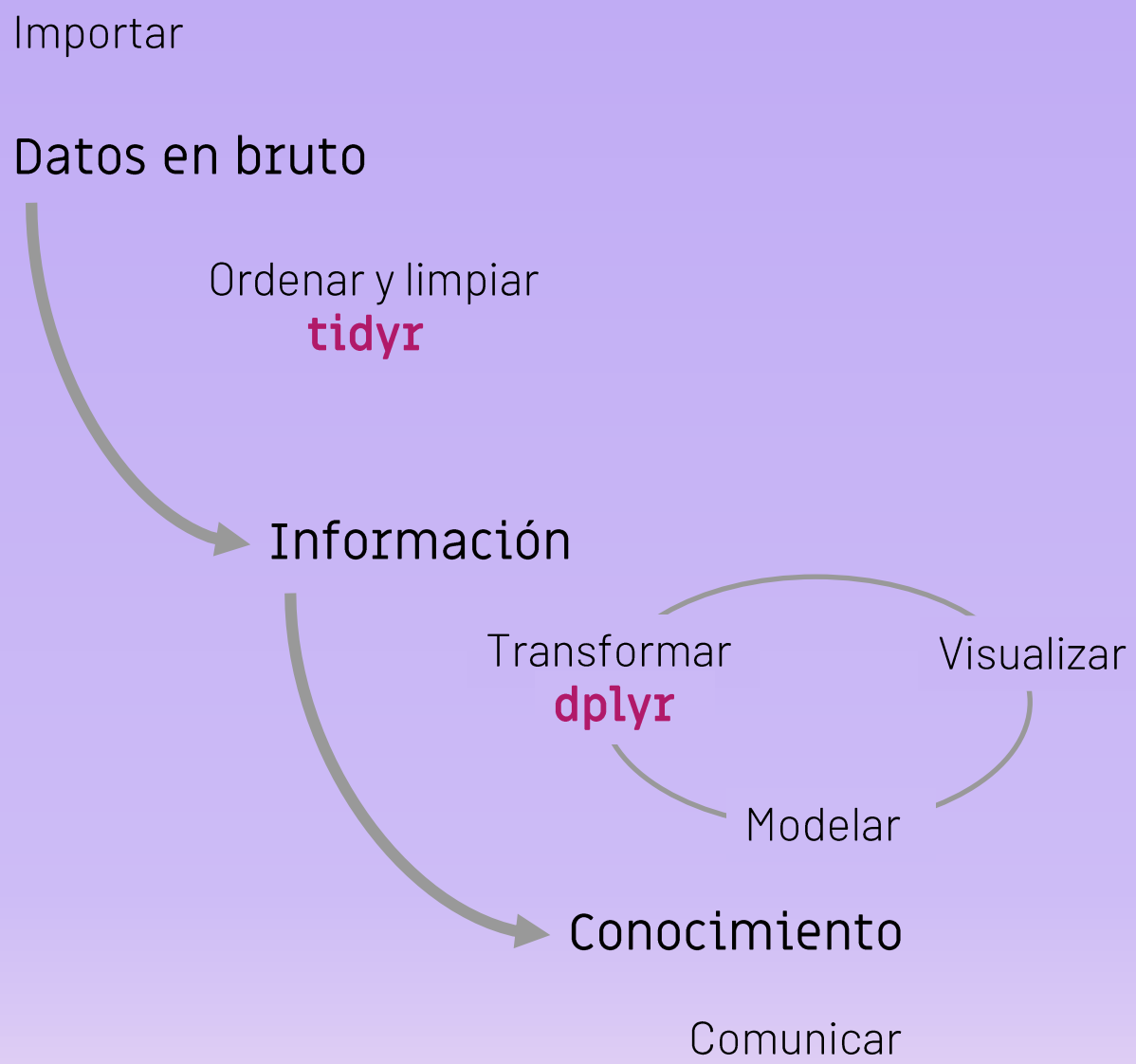
- Datos organizados y estructurados
- Presentados en un contexto
- Útiles para el que los observa

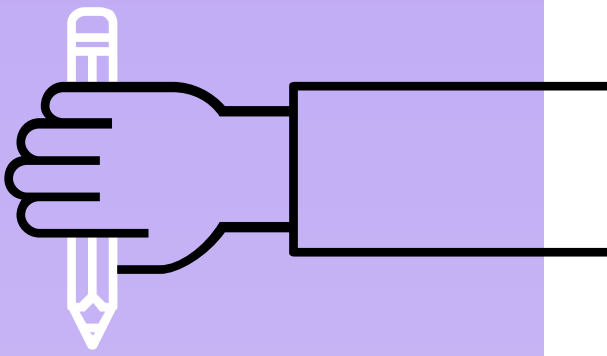
Conocimiento

- Procesar la información para extraer conclusiones valiosas
- *Modelación*
- Comunicación de resultados

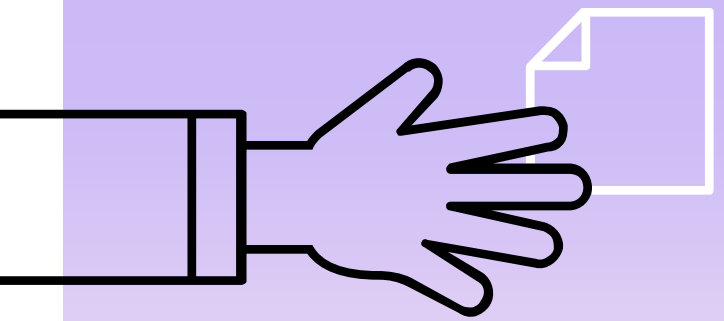
CIENCIA de datos

Roadmap del proceso en R





1. Ordenar



Estructurar las bases de datos a nuestra conveniencia para facilitar el análisis

Hadley Wickham.
Tidy data.
The Journal of Statistical Software, vol. 59, 2014

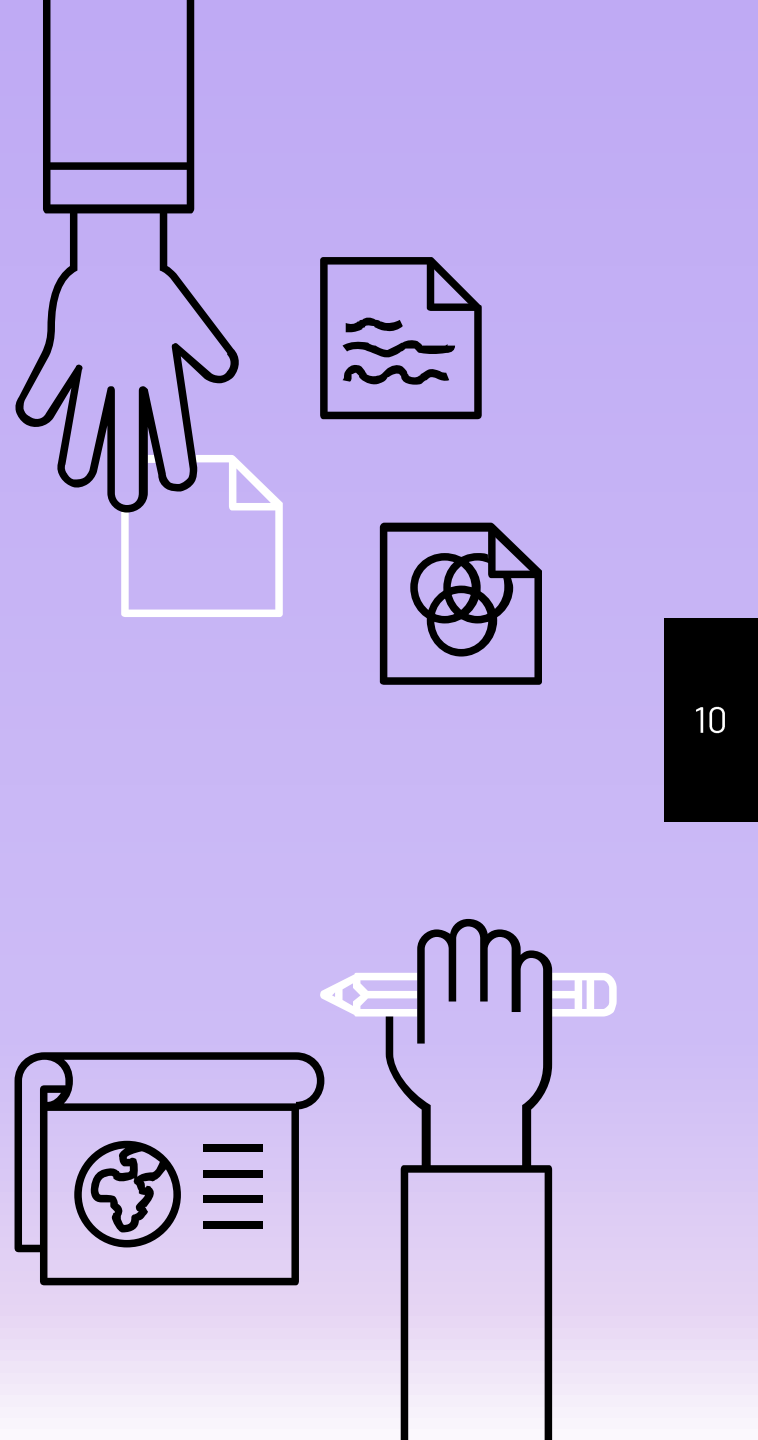
80%

Del tiempo invertido en un análisis
se lleva en limpiar y preparar los datos

Estándares y definición de datos ordenados (tidy data)

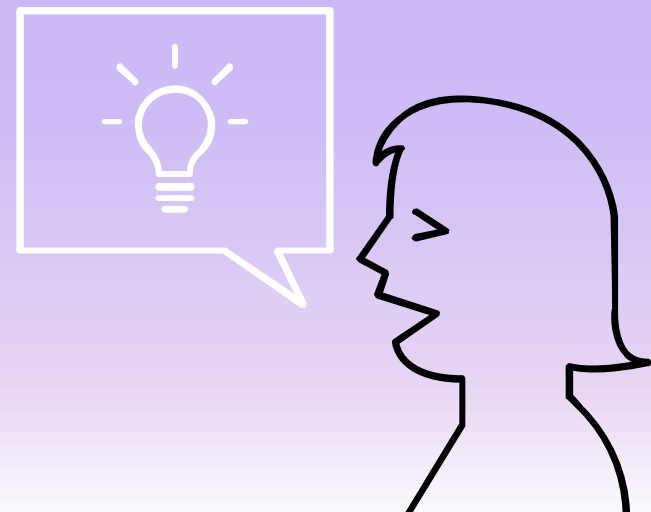
- Diseñados para estandarizar la forma de la base tal que nos facilite la exploración inicial y modelación

- Seguir los principios nos permitirá enfocarnos en el análisis del problema más que en la “logística” de los datos



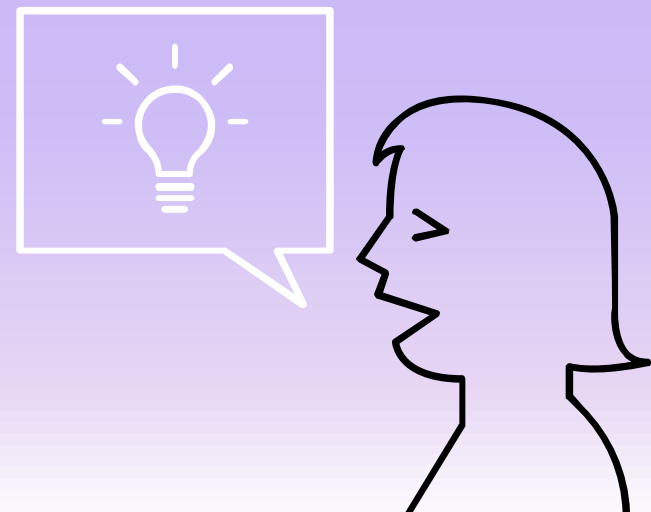
Todas las familias felices se parecen; cada familia infeliz lo es a su propia manera

— Leo Tolstoy



Todas las bases de datos ordenadas se parecen; pero cada base de datos desordenada lo es a su propia manera

— *Hadley Wickham*

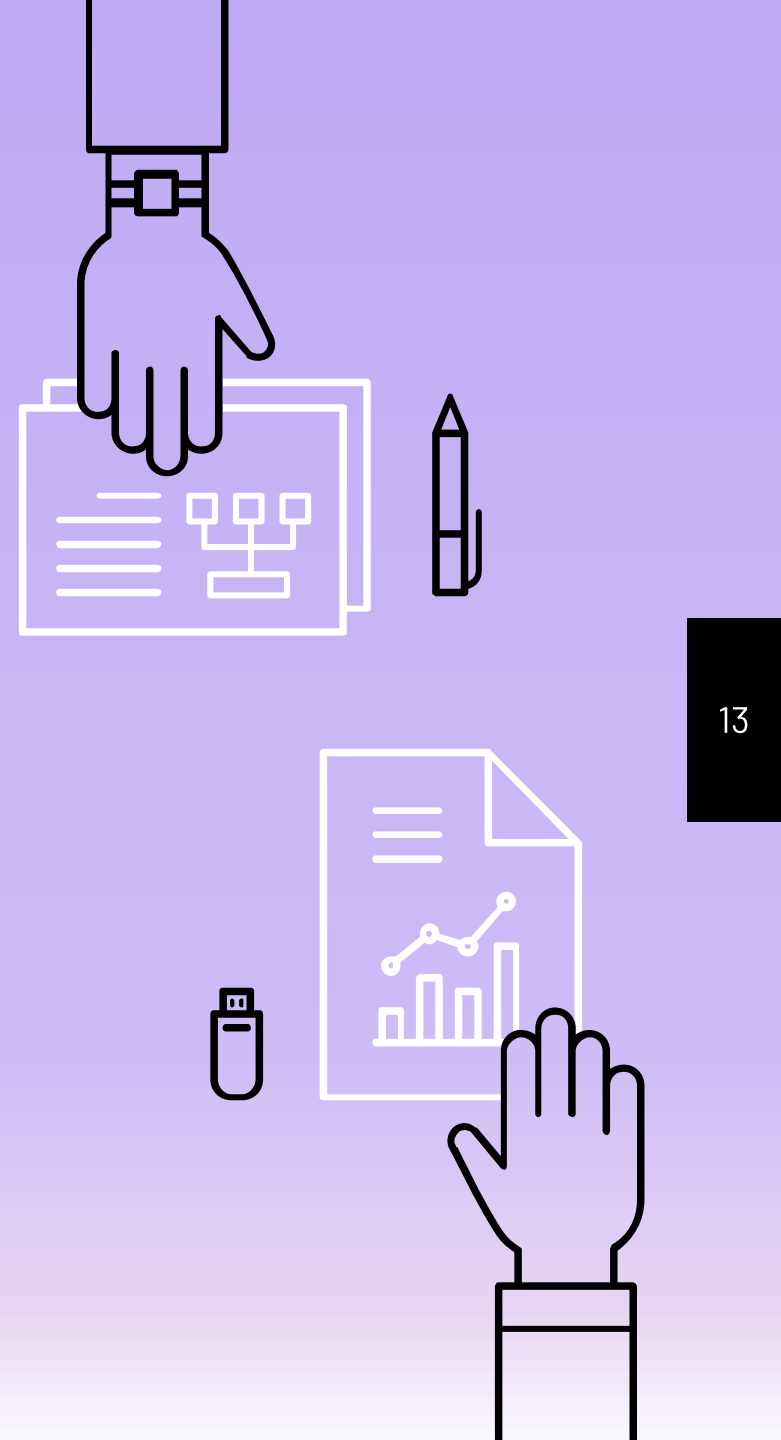


Estructura de una 'base de datos ordenada'

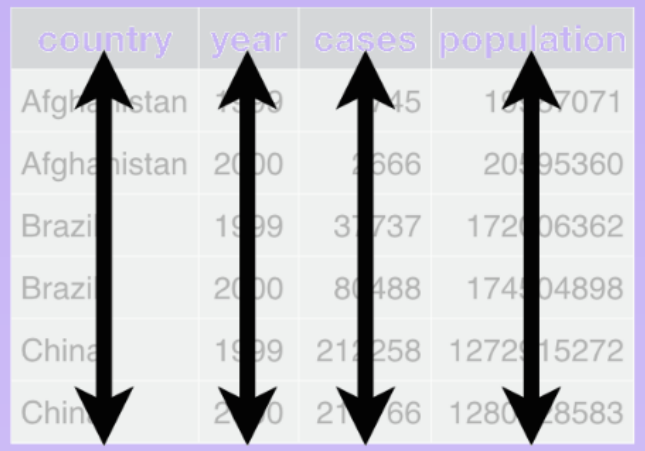
(Abrir script src/01_intro_tidyr.R)

Típicamente usaremos objetos de la clase data.frame, donde:

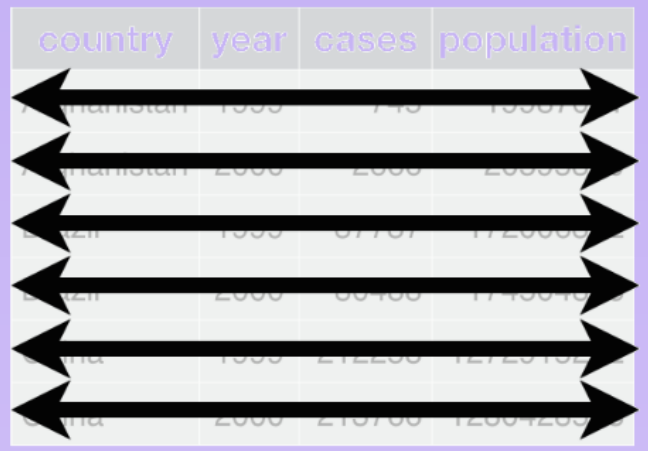
- Cada **celda** de la base corresponda a un valor (numérico, carácter, etc.) asociado a una variable y una observación
- Una **variable** (columna) contiene las mediciones de un mismo atributo en las unidades
- Una **observación** (fila) contiene las mediciones de todos los atributos sobre la misma unidad



Estructura de una ‘base de datos ordenada’



Variables



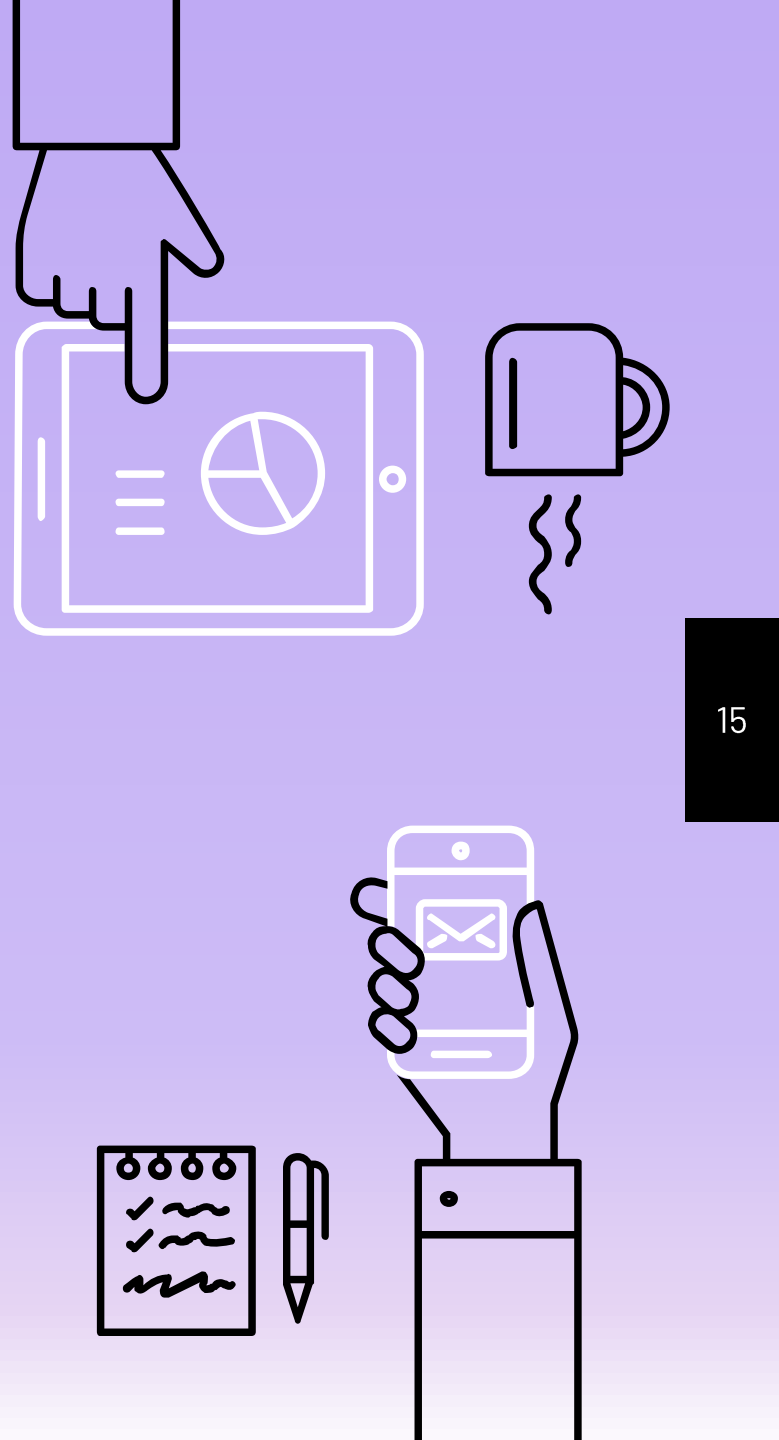
Observaciones

Valores

‘Datos ordenados’

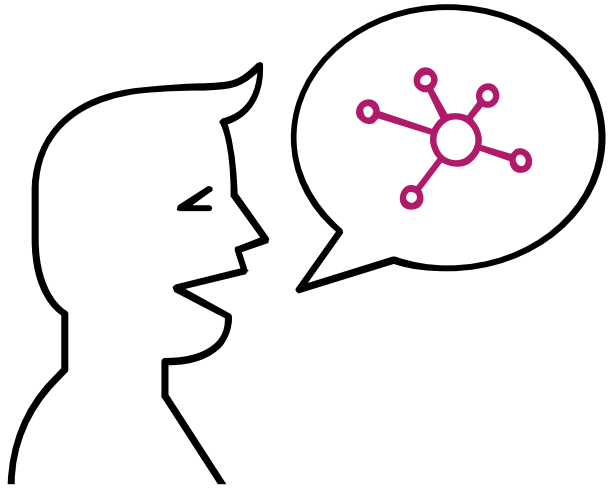
- Cada variable forma una columna
- Cada observación forma una fila
- Cada unidad de observación (personas, escuelas, países) forma una tabla

*Variables que son parte del diseño o que pueden usarse como moderadores categóricos deben colocarse al principio (izquierda)



Desorden #1

Los encabezados de columna son valores, no nombres de variables



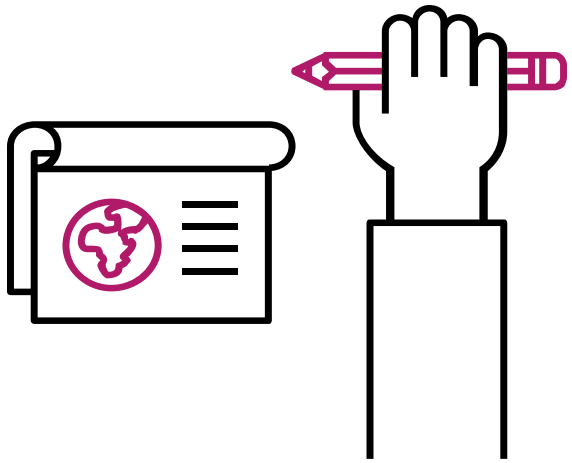
¿Cómo propondrían ordenarlo?

Desordenado

religion	<$10k	$10–20k	$20–30k	$30–40k	$40–50k	$50–75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

Desorden #1

Los encabezados de columna son valores, no nombres de variables



Ordenado

religion	income	freq
Agnostic	<$10k	27
Agnostic	$10–20k	34
Agnostic	$20–30k	60
Agnostic	$30–40k	81
Agnostic	$40–50k	76
Agnostic	$50–75k	137
Agnostic	$75–100k	122
Agnostic	$100–150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

Verbos para estructurar datos: gather

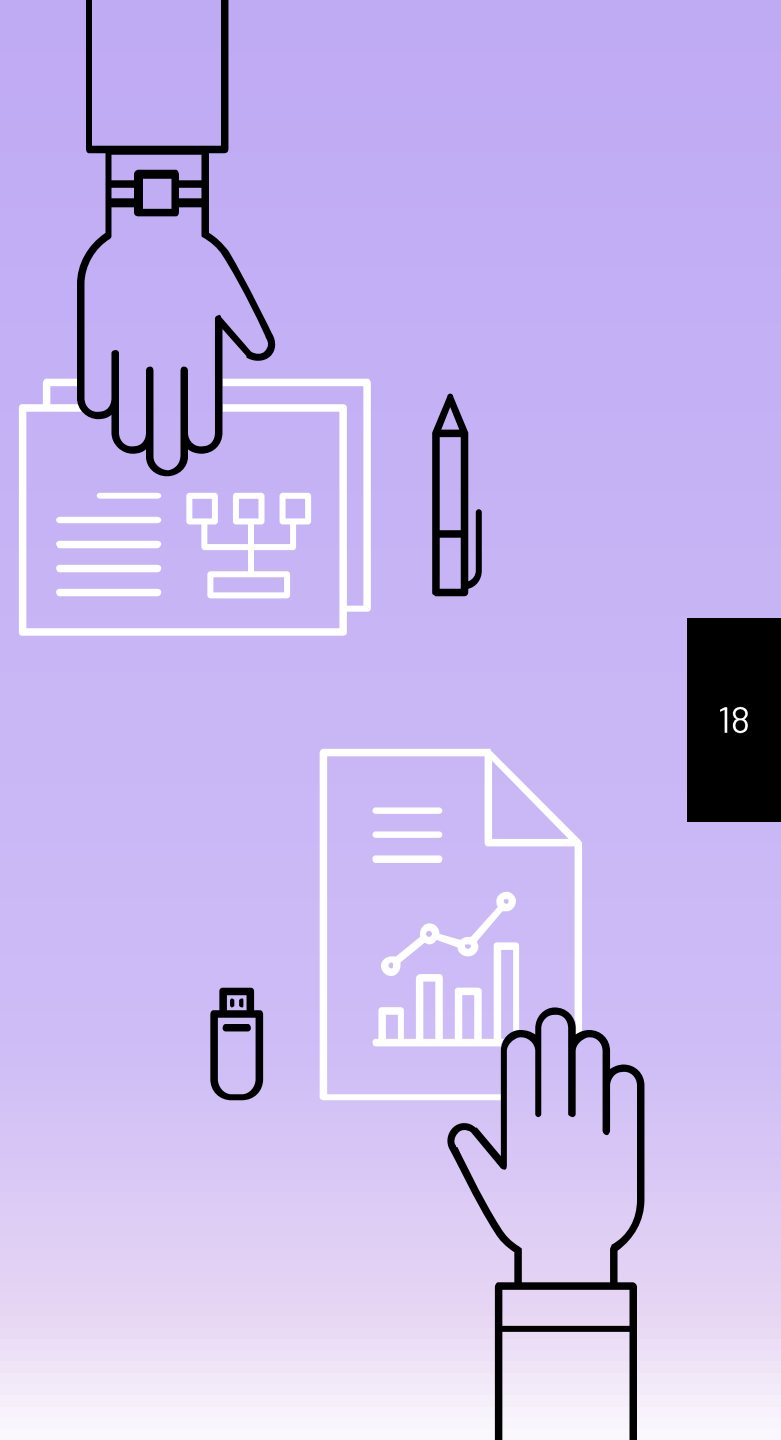
- Combina múltiples columnas en una sola, usando un formato de llave-valor

Desordenado

id	trt	work.T1	home.T1	work.T2	home.T2
1	treatment	0.08513597	0.6158293	0.1135090	0.05190332
2	control	0.22543662	0.4296715	0.5959253	0.26417767
3	treatment	0.27453052	0.6516557	0.3580500	0.39879073
4	control	0.27230507	0.5677378	0.4288094	0.83613414

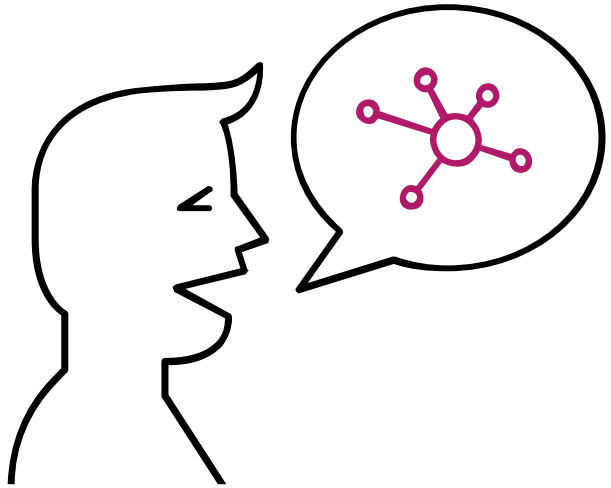
Más ordenado

id	trt	key	time
1	treatment	work.T1	0.08513597
2	control	work.T1	0.22543662
3	treatment	work.T1	0.27453052
4	control	work.T1	0.27230507
1	treatment	home.T1	0.61582931
2	control	home.T1	0.42967153
3	treatment	home.T1	0.65165567
4	control	home.T1	0.56773775
1	treatment	work.T2	0.11350898
2	control	work.T2	0.59592531
3	treatment	work.T2	0.35804998
4	control	work.T2	0.42880942
1	treatment	home.T2	0.05190332
2	control	home.T2	0.26417767
3	treatment	home.T2	0.39879073
4	control	home.T2	0.83613414



Desorden #2

Una columna contiene datos de más de una variable



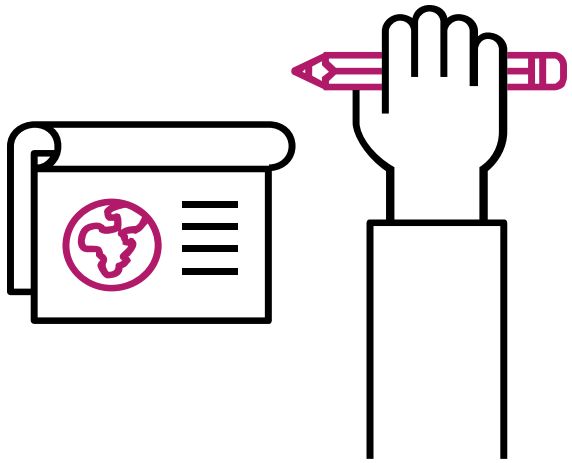
¿Cómo propondrían ordenarlo?

Desordenado

country	year	m014	m1524	m2534	m3544	m4554	m5564	m65	mu	f014
AD	2000	0	0	1	0	0	0	0	—	—
AE	2000	2	4	4	6	5	12	10	—	3
AF	2000	52	228	183	149	129	94	80	—	93
AG	2000	0	0	0	0	0	0	1	—	1
AL	2000	2	19	21	14	24	19	16	—	3
AM	2000	2	152	130	131	63	26	21	—	1
AN	2000	0	0	1	2	0	0	0	—	0
AO	2000	186	999	1003	912	482	312	194	—	247
AR	2000	97	278	594	402	419	368	330	—	121
AS	2000	—	—	—	—	1	1	—	—	—

Desorden #2

Una columna contiene datos de más de una variable



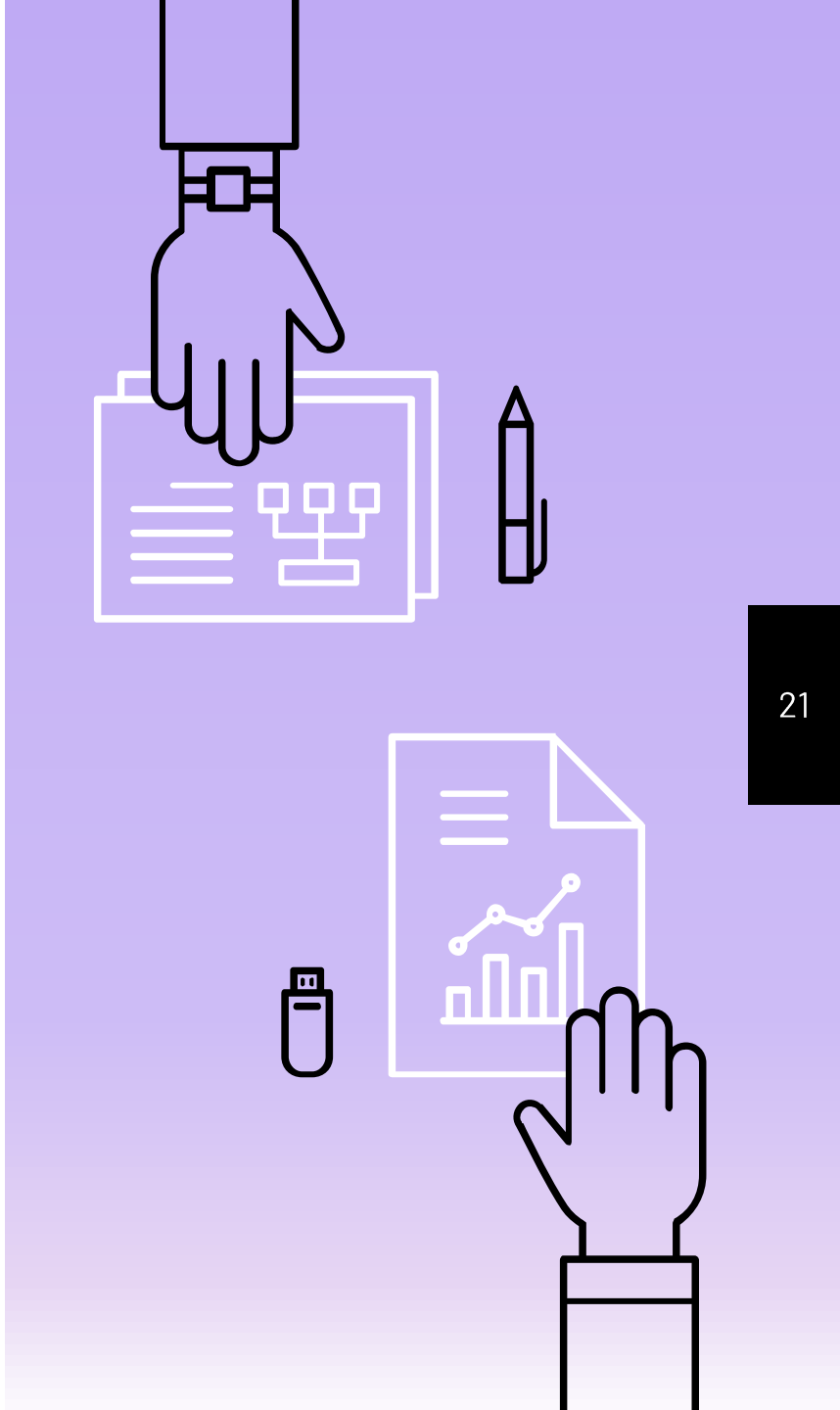
Ordenado

country	year	sex	age	cases
AD	2000	m	0–14	0
AD	2000	m	15–24	0
AD	2000	m	25–34	1
AD	2000	m	35–44	0
AD	2000	m	45–54	0
AD	2000	m	55–64	0
AD	2000	m	65+	0
AE	2000	m	0–14	2
AE	2000	m	15–24	4
AE	2000	m	25–34	4
AE	2000	m	35–44	6
AE	2000	m	45–54	5
AE	2000	m	55–64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3

Verbos para estructurar datos: separate

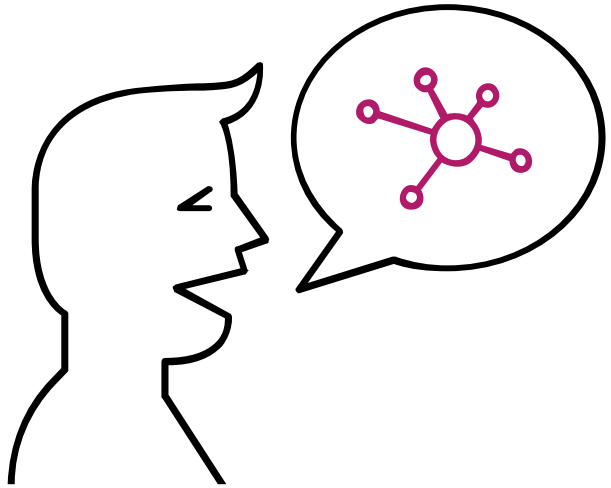
- Separa una variable en múltiples variables
- Útil cuando un valor representa más de un atributo (por ejemplo sexo y edad)

separate(data, col, into, sep, remove = TRUE)



Desorden #3

Variables tanto en filas como en columnas



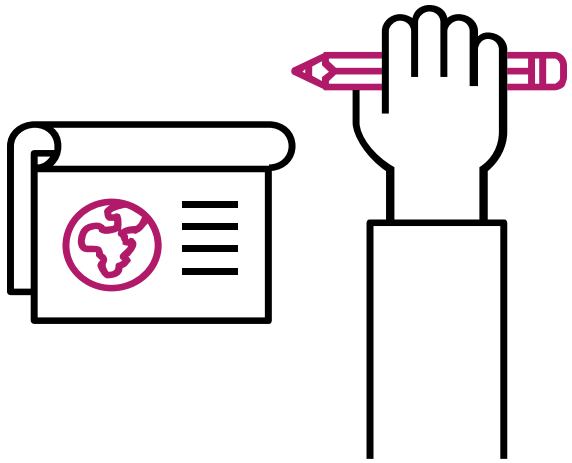
¿Cómo propondrían ordenarlo?

Desordenado

id	year	month	element	d1	d2	d3	d4	d5	d6	d7	d8
MX17004	2010	1	tmax	—	—	—	—	—	—	—	—
MX17004	2010	1	tmin	—	—	—	—	—	—	—	—
MX17004	2010	2	tmax	—	27.3	24.1	—	—	—	—	—
MX17004	2010	2	tmin	—	14.4	14.4	—	—	—	—	—
MX17004	2010	3	tmax	—	—	—	—	32.1	—	—	—
MX17004	2010	3	tmin	—	—	—	—	14.2	—	—	—
MX17004	2010	4	tmax	—	—	—	—	—	—	—	—
MX17004	2010	4	tmin	—	—	—	—	—	—	—	—
MX17004	2010	5	tmax	—	—	—	—	—	—	—	—
MX17004	2010	5	tmin	—	—	—	—	—	—	—	—

Desorden #3

Variables tanto en filas como en columnas



'Gathereado'

	id	year	month	day	element	value
1	MX17004	2010	1	30.00	tmax	27.80
2	MX17004	2010	1	30.00	tmin	14.50
3	MX17004	2010	2	2.00	tmax	27.30
4	MX17004	2010	2	2.00	tmin	14.40
5	MX17004	2010	2	3.00	tmax	24.10
6	MX17004	2010	2	3.00	tmin	14.40
7	MX17004	2010	2	11.00	tmax	29.70
8	MX17004	2010	2	11.00	tmin	13.40
9	MX17004	2010	2	23.00	tmax	29.90
10	MX17004	2010	2	23.00	tmin	10.70
11	MX17004	2010	3	5.00	tmax	32.10
12	MX17004	2010	3	5.00	tmin	14.20
13	MX17004	2010	3	10.00	tmax	34.50
14	MX17004	2010	3	10.00	tmin	16.80
15	MX17004	2010	3	16.00	tmax	31.10
16	MX17004	2010	3	16.00	tmin	17.60
17	MX17004	2010	4	27.00	tmax	36.30
18	MX17004	2010	4	27.00	tmin	16.70
19	MX17004	2010	5	27.00	tmax	33.20
20	MX17004	2010	5	27.00	tmin	18.20

Desorden #3

Variables tanto en filas
como en columnas

Verbos para estructurar datos: spread

- Divide renglones llave-valor en columnas

Ordenado

id	date	tmax	tmin
MX17004	2010-01-30	27.8	14.5
MX17004	2010-02-02	27.3	14.4
MX17004	2010-02-03	24.1	14.4
MX17004	2010-02-11	29.7	13.4
MX17004	2010-02-23	29.9	10.7
MX17004	2010-03-05	32.1	14.2
MX17004	2010-03-10	34.5	16.8
MX17004	2010-03-16	31.1	17.6
MX17004	2010-04-27	36.3	16.7
MX17004	2010-05-27	33.2	18.2

Resumen: verbos para estructurar datos

gather

Combina múltiples columnas en una sola regresando un formato llave-valor

spread

Divide filas llave-valor en columnas

unite

Combina dos columnas (variables) en una sola, pegándolas

separate

Separa una variable en más de una. Es muy útil cuando los valores representan más de un atributo