Limpieza semiautomática de datos con el paquete clickR

David Hervás Marín – Universitat Politècnica de València

Valencia 05/11/2025

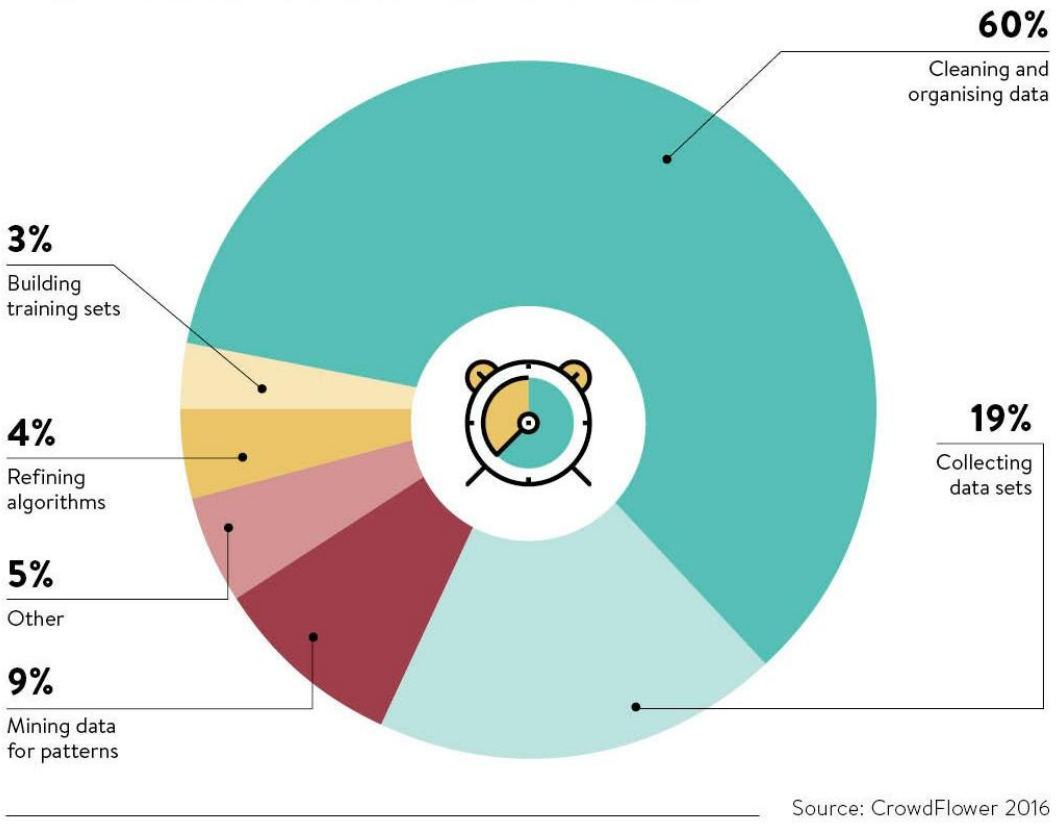
UNIVERSITAT POLITÈCNICA DE VALÈNCIA

LIMPIEZA DE DATOS

When I started cleaning data

When I finished cleaning data

LIMPIEZA DE DATOS



ALTERNATIVAS

Limpieza manual

- ✓ Control detallado del resultado
- ✓ Facilita la comprensión de los datos
- ✗ Consume mucho tiempo
- ✗ Sujeta a fallos humanos

Limpieza automática

- ✓ Rápida
- ✓ Puede ser desatendida
- ✗ Poco control de los resultados
- ✗ Los algoritmos/la IA no funcionan bien en todos los casos

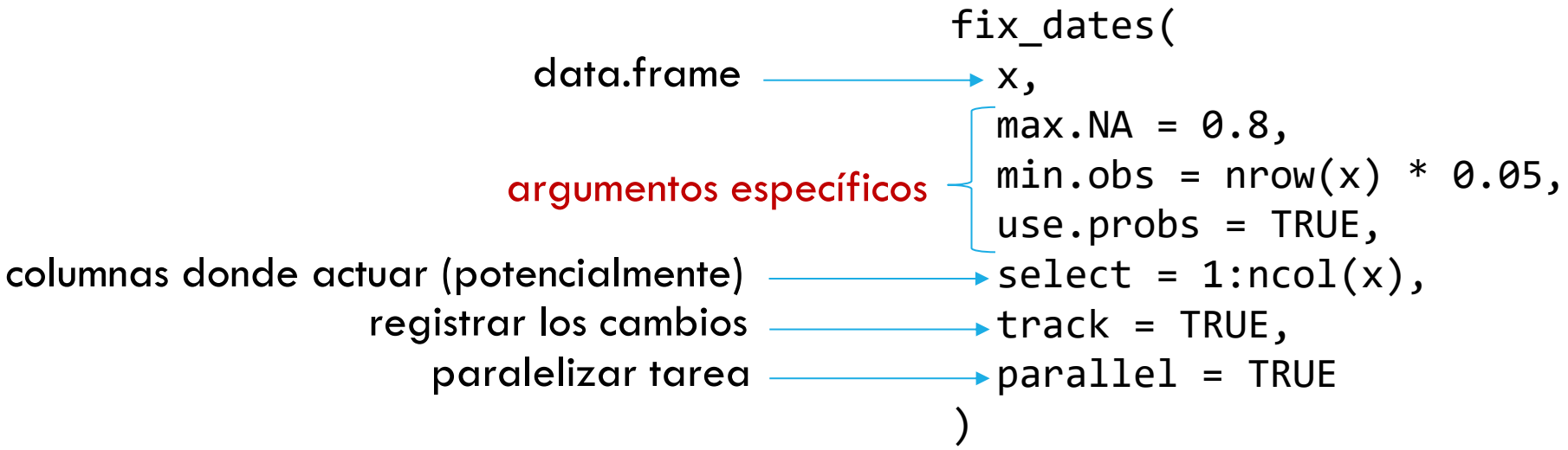
LIMPIEZA SEMIAUTOMÁTICA

El software propone e implementa correcciones y nosotros las revisamos y validamos (o revertimos).

- Herramientas de análisis exploratorio
- Funciones para detección y corrección de errores
- Registro de cambios con función de “deshacer”
- Funciones miscelánea: `f_replace()`, `good2go()`, etc.

ESTRUCTURA DE LAS FUNCIONES DE LIMPIEZA

Casi todas las funciones 'fix_' tienen la misma estructura:



FUNCIONAMIENTO

Todas las funciones 'fix_' admiten un data.frame como entrada y devuelven ese data.frame corregido como salida. En el data.frame de salida se guarda (opcionalmente) el registro de cambios como un atributo, que también es un data.frame.

```
datos <- fix_dates(datos)
attr(datos, "changes")
```

```
track_changes(datos, subset)
```

variable	observation	original	new	fun
x1	all	character	Date	fix_dates
x2	all	character	Date	fix_dates
x1	1	25/06/1983	1983-06-25	fix_dates
x1	2	25-08/2014	2014-08-25	fix_dates
x1	3	2001/11/01	2001-11-01	fix_dates
x2	1	01/01/85	1985-01-01	fix_dates
x2	2	04/04/1982	1982-04-04	fix_dates
x2	3	07/12-2016	2016-12-07	fix_dates
x2	4	September 24, 2020	2020-09-24	fix_dates

¿QUÉ FUNCIONES HAY DISPONIBLES EN CLICKR?

Funciones para detección y corrección de errores

```
"fix_concat"        "fix_dates"        "fix_factors"      "remove_empty"    "manual_fix"
"fix_levels"        "fix_NA"           "fix_numerics"     "nice_names"      "restore_changes“
"track_changes"
```

Funciones para análisis exploratorio

```
"bivariate_outliers" "check_quality"     "cluster_var"       "descriptive“    "GK_assoc“
"mine.plot"             "ipboxplot"        "outliers"              "peek"
```

Funciones misceláneas

```
"%<=NA%"            "%<NA%"            "%>=NA%"            "%>NA%"         "%between%"
"%betweenNA%"       "f_replace"        "good2go"
```

EJEMPLO SENCILLO: `mtcars_messy`

El paquete *clickR* incluye una versión "sucia" del conocido data set "mtcars":

	Mpg	cyl	disp	hp	drat	wt	qsec	vs	am	n Gears	carb	date	maker
Datsun 710	22,8	4	108.0	93	3.85	2.32	18.61	1	1	4	1	1973-12-23	Datsun
Hornet 4 Drive	21.,4	6	258.0	110	3.08	3.215	19.44	1	0	3	1	1974-06-16	ornet
Duster 360	14.3 mpg	8	360.0	245		3.57	15.84	0	0	3	4	26/08/1974	Duster
Merc 280C	1.78e01	6	167.6	123	3.92	3.44	18.90	1	0	4	4	3rd July 1974	Merc
Lincoln Continental	10.4	8	460.0	215	3	5,424	17.82	0	0	3	4	1973-12-17	Lincoln
Chrysler Imperial	14.7	8	440.0	230	3.23	5,345	17.42	0	0	3	4	1973-01-11	Chrysler
Honda Civic	30.4	4	75.7	52	-	1.615	18.52	1	1	4	2	1974-02-18	Honda
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1	19/06/74	Toyota
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2	1974-03-30	Pontiac
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1	1973-09-16	Fiat
Porsche 914-2	26	4	120.3	91	4.43	2.14	16.70	NULL	1	5	2	1973-07-26	Porsche
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2	14 /08-1974	Lotus
Ford Pantera L	15.8	8	351.0	264	4.22	3.17	14.50	0	1	5	4	1974-03-07	Ford
Ferrari Dino	19.7	6	145.0	175	3.62	2.77	15.50	0	1	5	6	1974-05-16	Ferrari
Maserati Bora	15	8	301.0	335	3.54	3.57	14.60	0	1	5	8	1973, 03. 03	Maserati
Volvo 142E	21.4	4	121.0	109	4.11	2.78	18.60	1	1	4	2	12/22/73	Volvo

¡VÁMONOS A R!

https://github.com/David-Hervas/clickR-tutorial