

# Manipulación de datos en Pandas



GOBIERNO  
DE ESPAÑA

VICEPRESIDENCIA  
PRIMERA DEL GOBIERNO  
MINISTERIO  
DE ASUNTOS ECONÓMICOS  
Y TRANSFORMACIÓN DIGITAL

SECRETARÍA DE ESTADO  
DE DIGITALIZACIÓN  
E INTELIGENCIA ARTIFICIAL

red.es

Centro de  
Referencia Nacional  
en Comercio Electrónico  
y Marketing

CRN  
Digital



UNIÓN EUROPEA

Barrabés

The Valley

"El FSE invierte en tu futuro"  
Fondo Social Europeo

1. CRISP-DM
2. EDA
3. Definiciones importantes
  - 2.1 Observaciones y características
  - 2.2 Missing y outliers
  - 2.3 Data cleaning y data wrangling
4. Variables
  - 3.1 Numéricas
  - 3.2 Categóricas
  - 3.3 Tipos de datos
5. Data preparation
6. Pandas
7. Visualización



# CRISP-DM



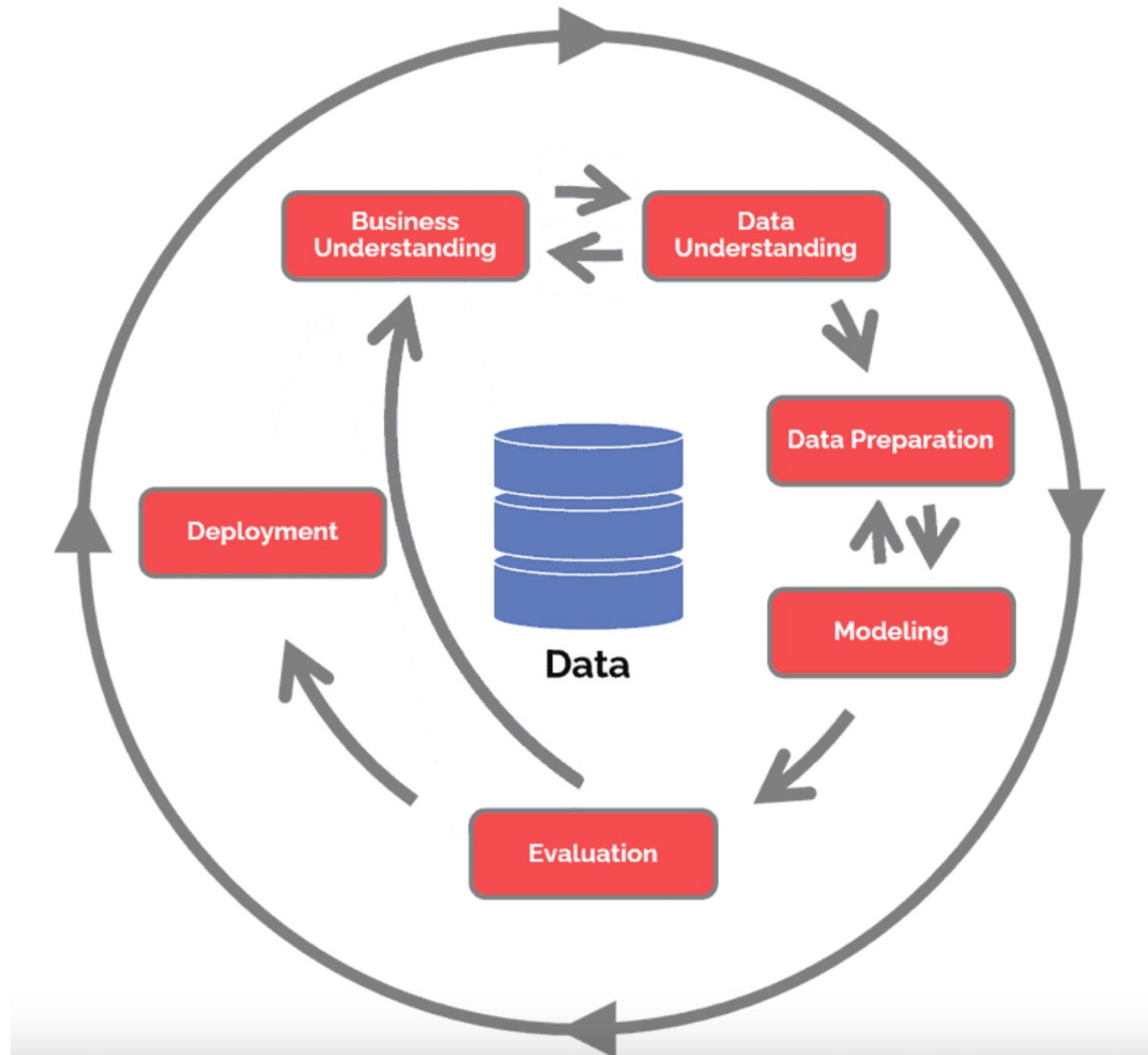
# CRISP-DM

## ¿Qué es el CRISP-DM?

**Por sus siglas en inglés, Cross Industry Standard Process for Data Mining.**

**Es un estándar que sirve como modelo en el proceso de ciencias de datos.**

## CRISP-DM



# EDA



# EDA

## ¿Qué es el EDA?

**Por sus siglas en inglés, Exploratory Data Analysis, es el análisis exploratorio de datos.**

**Una de las tareas principales de un científico de datos.**

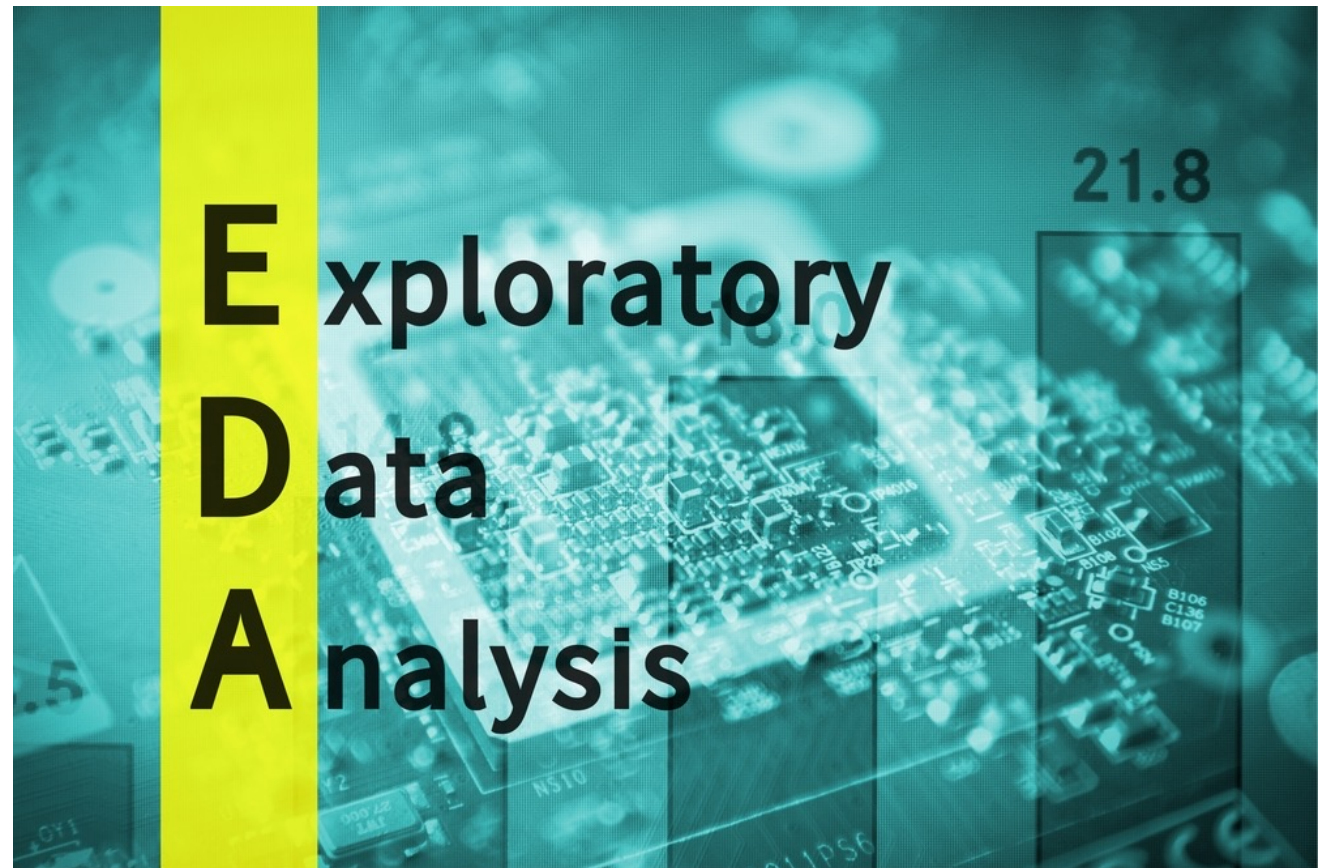
**Trata de buscar patrones y distribuciones en los datos, así como detectar anomalías y comprobar suposiciones con la ayuda de resúmenes estadísticos y representaciones gráficas.**



# EDA

## Nos permitirá

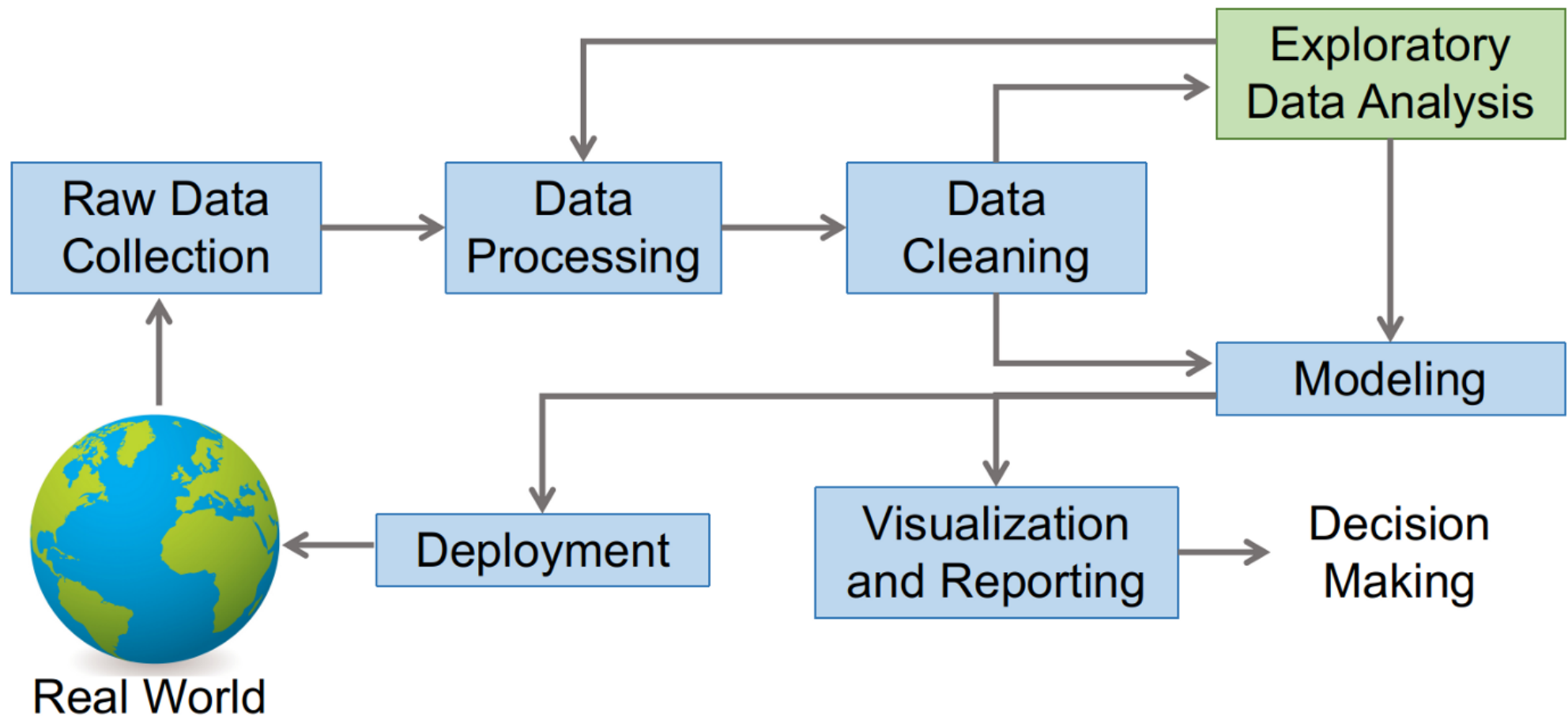
1. Conocer los datos
2. Identificar patrones
3. Detectar *outliers*



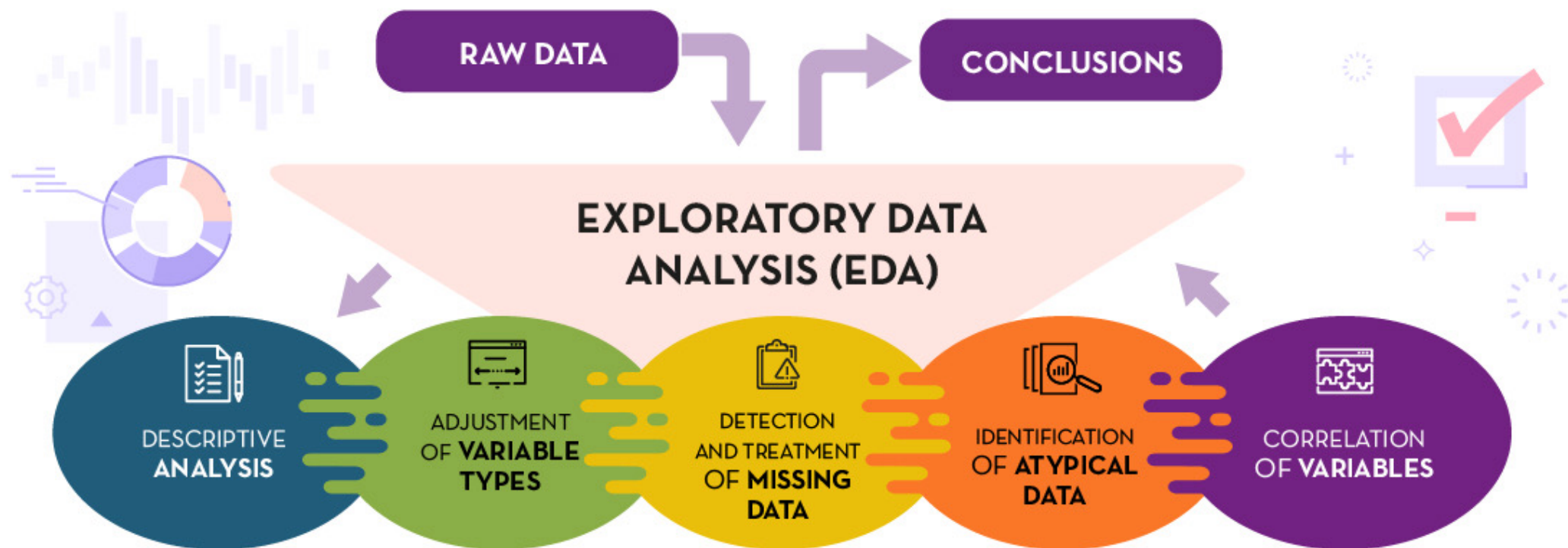


## EDA

### Data Science Process



# EDA



# Definiciones importantes



## Definiciones importantes

### Observaciones

También se les denomina **registros**.

Hace referencia a cada una de las **filas** de la base de datos.

### Características

También se les denomina ***features***.

Hace referencia a cada una de las **columnas** en una base de datos.

## Definiciones importantes

### *Missing data*

Son los valores perdidos de la base de datos.

Se les puede llamar **nulos**.

A veces son campos vacíos, NaN, None, 0, -1, ...

### *Outliers*

Son observaciones de la base de datos que se alejan de la distribución del resto. Es decir, son muestras muy diferentes a las demás.

## Definiciones importantes

### DATA WRANGLING VERSUS DATA CLEANING

#### DATA WRANGLING

Process of transforming and mapping data from one raw data form into another form with the intent of making it more appropriate and valuable for various tasks

Data munging is another name for data wrangling

#### DATA CLEANING

Process of detecting and removing corrupted or inaccurate records from a record set, table or database

Data cleansing is another name for data cleaning

Visit [www.PEDIAA.com](http://www.PEDIAA.com)

## Preguntas ante una base de datos

- ¿Cuántos registros hay?
- ¿Son demasiado pocos?
- ¿Son muchos y no tenemos capacidad (CPU+RAM) suficiente para procesarlos?
- ¿Están todas las filas completas o tenemos campos con valores nulos?
- En caso de que haya demasiados valores nulos, ¿queda el resto de información inútil?
- ¿Cuáles parecen ser características importantes? ¿Cuáles podemos descartar?
- ¿Hay correlación entre características?

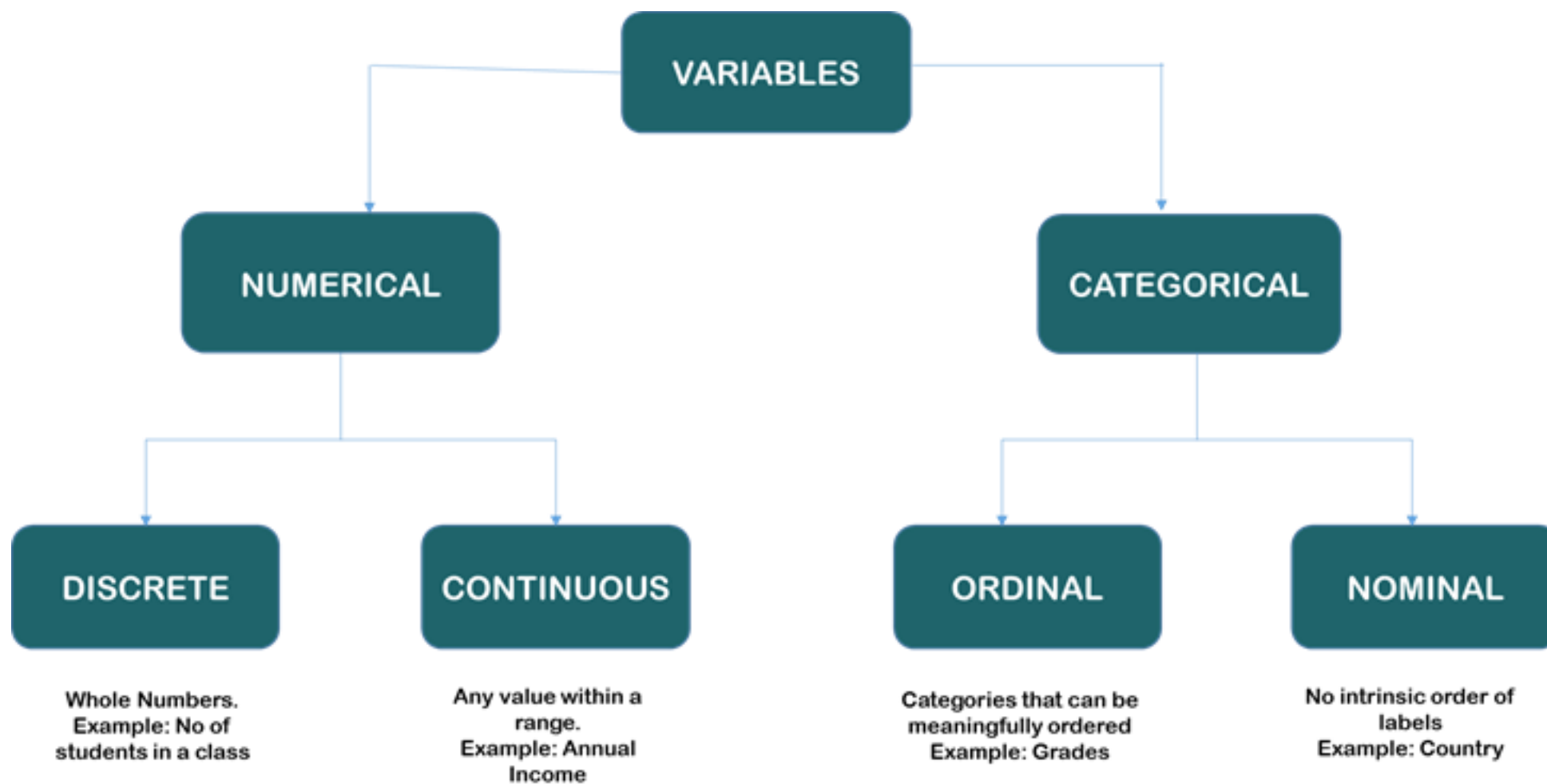




# Variables



# Variables



# Variables

## Tipos de datos

**Obtener el tipo de datos:**

- **Int**
- **Float**
- **Double**
- **String**
- **Bool**

## Preguntas ante una base de datos

- ¿Qué datos son discretos?
- ¿Cuáles son continuos?
- ¿Siguen alguna distribución?



# Data preparation



# Data preparation

## Proceso

**Una vez que tenemos los datos explorados....**

**Empezamos a detectar:**

- **Nulos (missing)**
- **Valores atípicos (Outliers)**
- **Datos incompletos**
- **Datos erróneos**

## Data preparation

### Nulos



Debemos detectarlos y colocar una señal que identifiquemos en nuestro código como *nulos*.

Unificar todos los nulos de la misma forma.

Student id	Marks in Maths (out of 100)	Marks in Maths (out of 100)	Remarks
19060641015	45	68	Good
19060641016	53	53	Bad
19060641017	68	78	Good
19060641018	75	75	Good
19060641019	80	45	Poor
19060641020	82	82	Good
19060641021	49	NULL	<u>NaN</u>
19060641022	76	80	Good
19060641023	79	79	Good
19060641024	55	55	Average
19060641025	80	52	Bad
19060641026	N/A	NULL	<u>NaN</u>
19060641027	N/A	87	<u>NaN</u>
19060641028	N/A	NULL	<u>NaN</u>

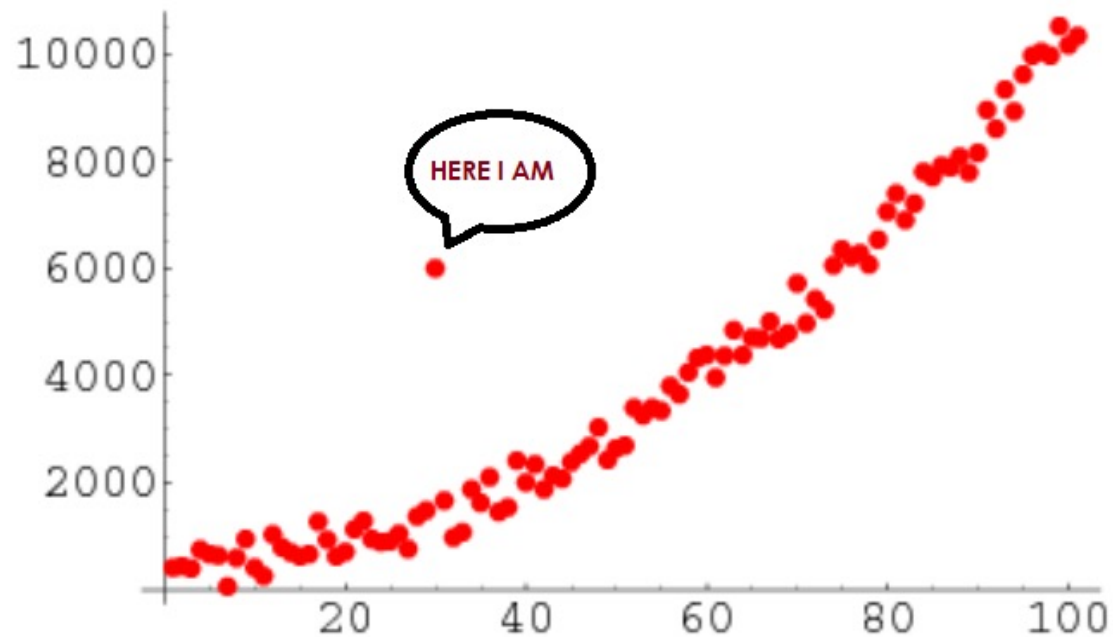
[NB: A dataset is a collection data points. Like age, weight etc.]



# Data preparation

## Outliers

Valores fuera de la norma general / de la distribución de los datos.



## Data preparation

### Incompletos

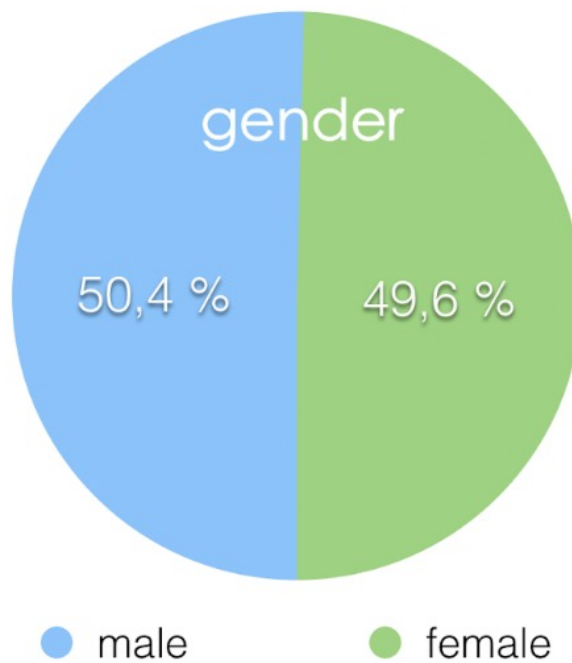
Hay veces que aún teniendo todo el histórico de los datos, no es suficientemente buena la calidad de los mismos y no es representativa la distribución que tenemos.

### Datos no balanceados

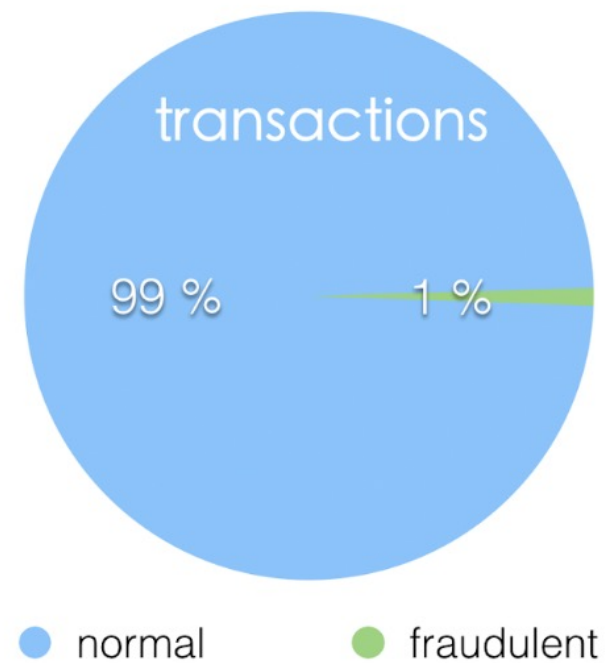
# Data preparation

## Incompletos

Balanced Dataset



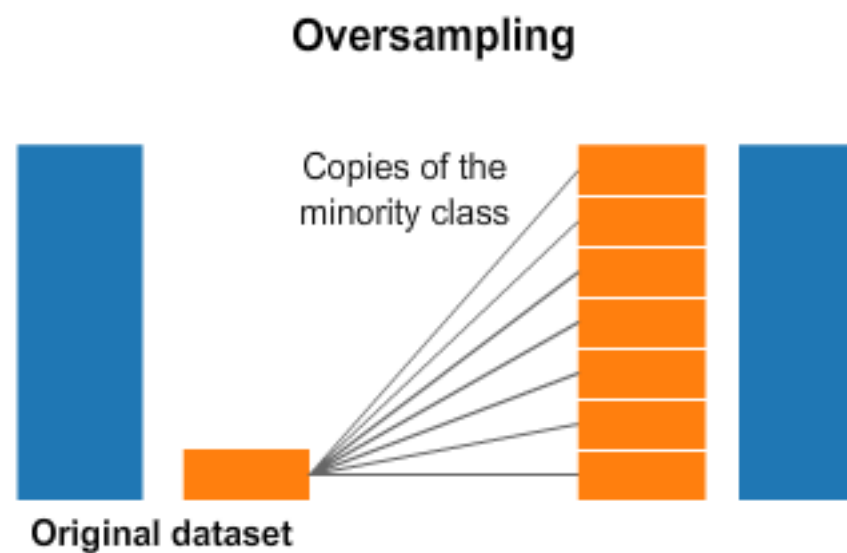
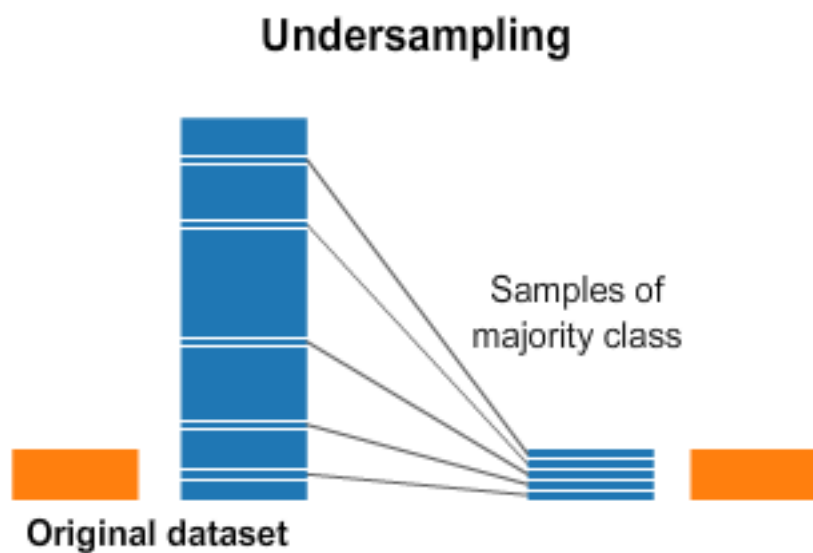
Unbalanced Dataset



## Data preparation

### Incompletos

¿Qué hacemos?



# Data preparation

## Erróneos

Una mala calidad en los datos también se ve reflejada en errores en las bases de datos.

- ¿Cómo son los datos?
- ¿Cuál es el rango de valores posible?
- ¿Existe un orden?



## Data preparation

**BAD DATA IS NO  
BETTER THAN NO  
DATA.**



## Data preparation





# Ejercicios



## Notebook 1

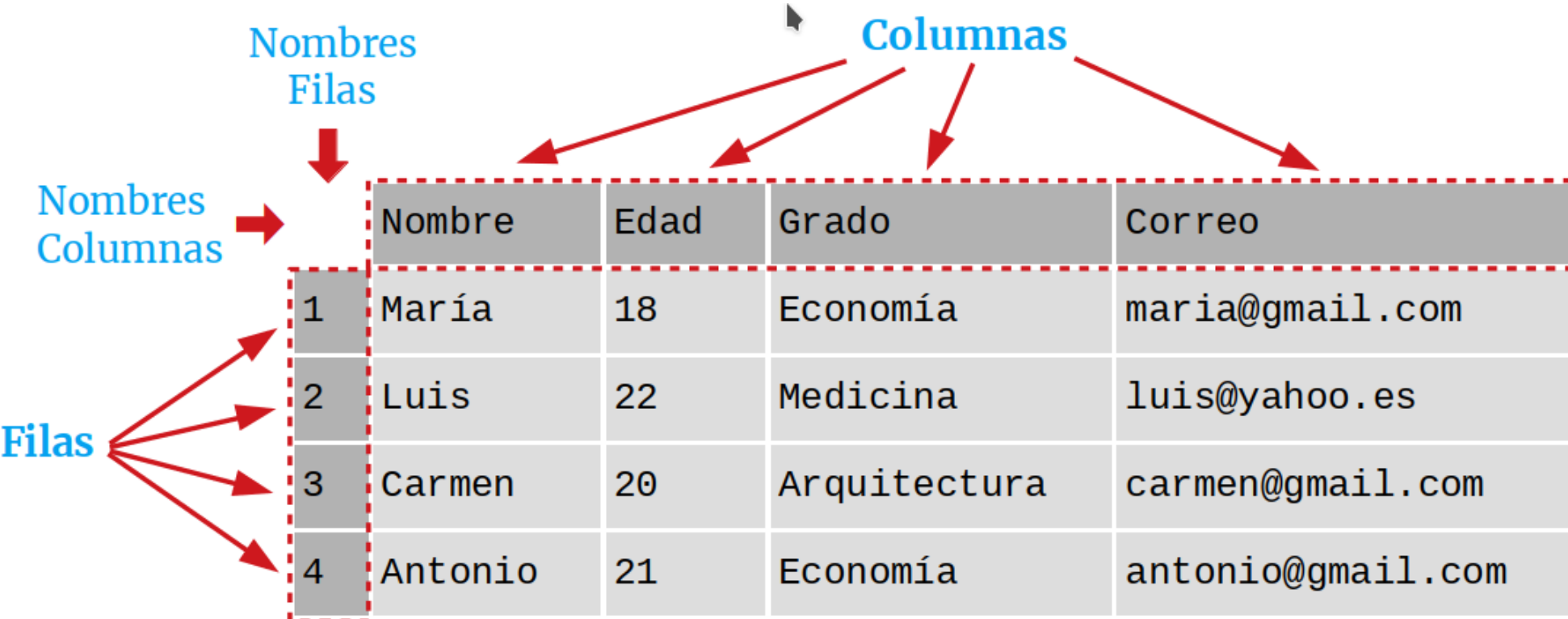


Manipulación de datos en Pandas

# Pandas



# Pandas

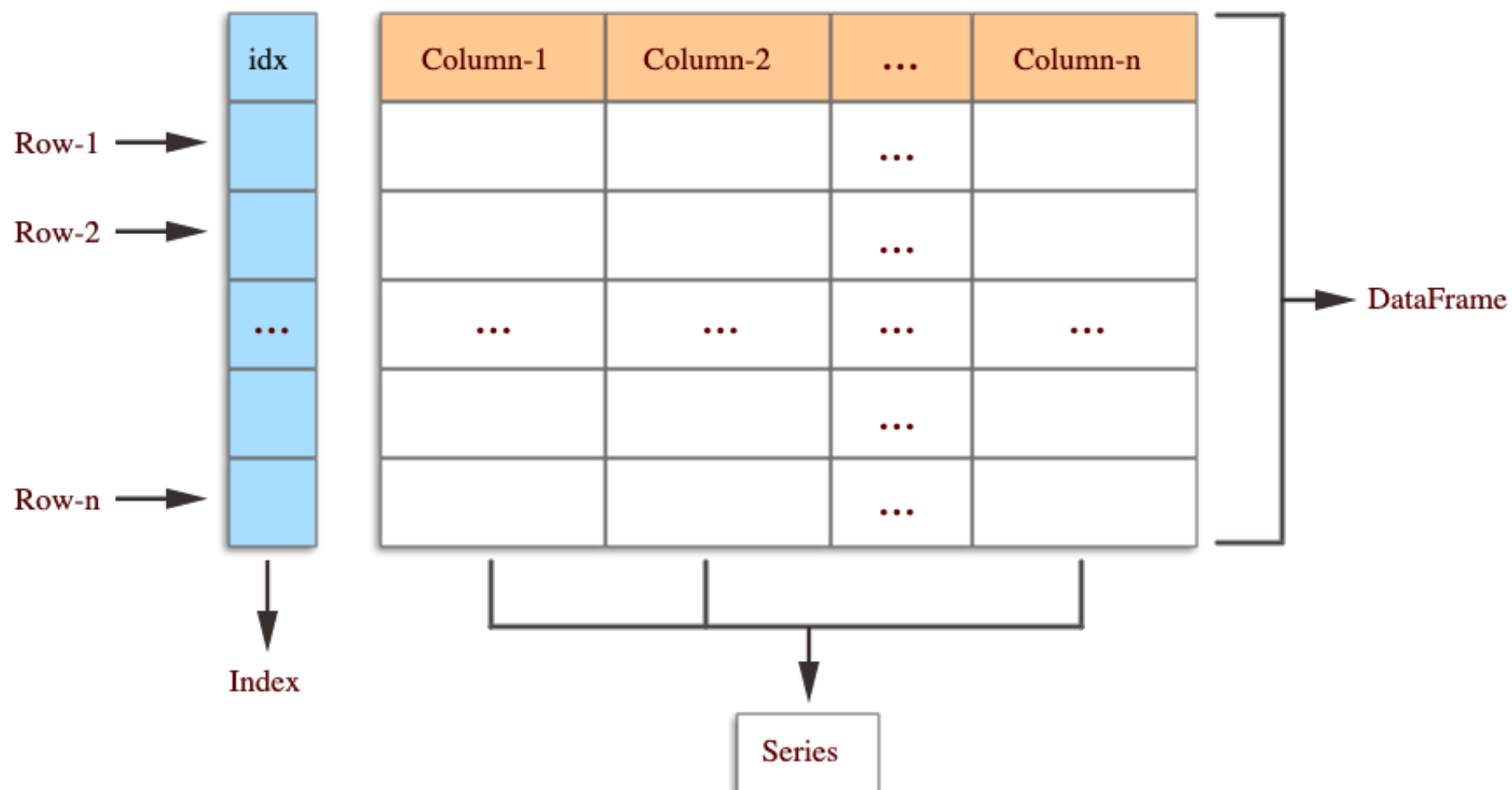


The diagram illustrates the structure of a Pandas DataFrame. It shows a table with 4 rows and 4 columns. The columns are labeled 'Nombre', 'Edad', 'Grado', and 'Correo'. The rows are indexed 1, 2, 3, and 4. The first row contains the names 'María', '18', 'Economía', and 'maria@gmail.com'. The second row contains 'Luis', '22', 'Medicina', and 'luis@yahoo.es'. The third row contains 'Carmen', '20', 'Arquitectura', and 'carmen@gmail.com'. The fourth row contains 'Antonio', '21', 'Economía', and 'antonio@gmail.com'. Red arrows point from the labels 'Nombres Filas' (Rows) and 'Nombres Columnas' (Columns) to the respective row and column headers. Red arrows also point from the label 'Filas' (Rows) to the row indices. Red arrows point from the label 'Columnas' (Columns) to the column headers.

	Nombre	Edad	Grado	Correo
1	María	18	Economía	maria@gmail.com
2	Luis	22	Medicina	luis@yahoo.es
3	Carmen	20	Arquitectura	carmen@gmail.com
4	Antonio	21	Economía	antonio@gmail.com

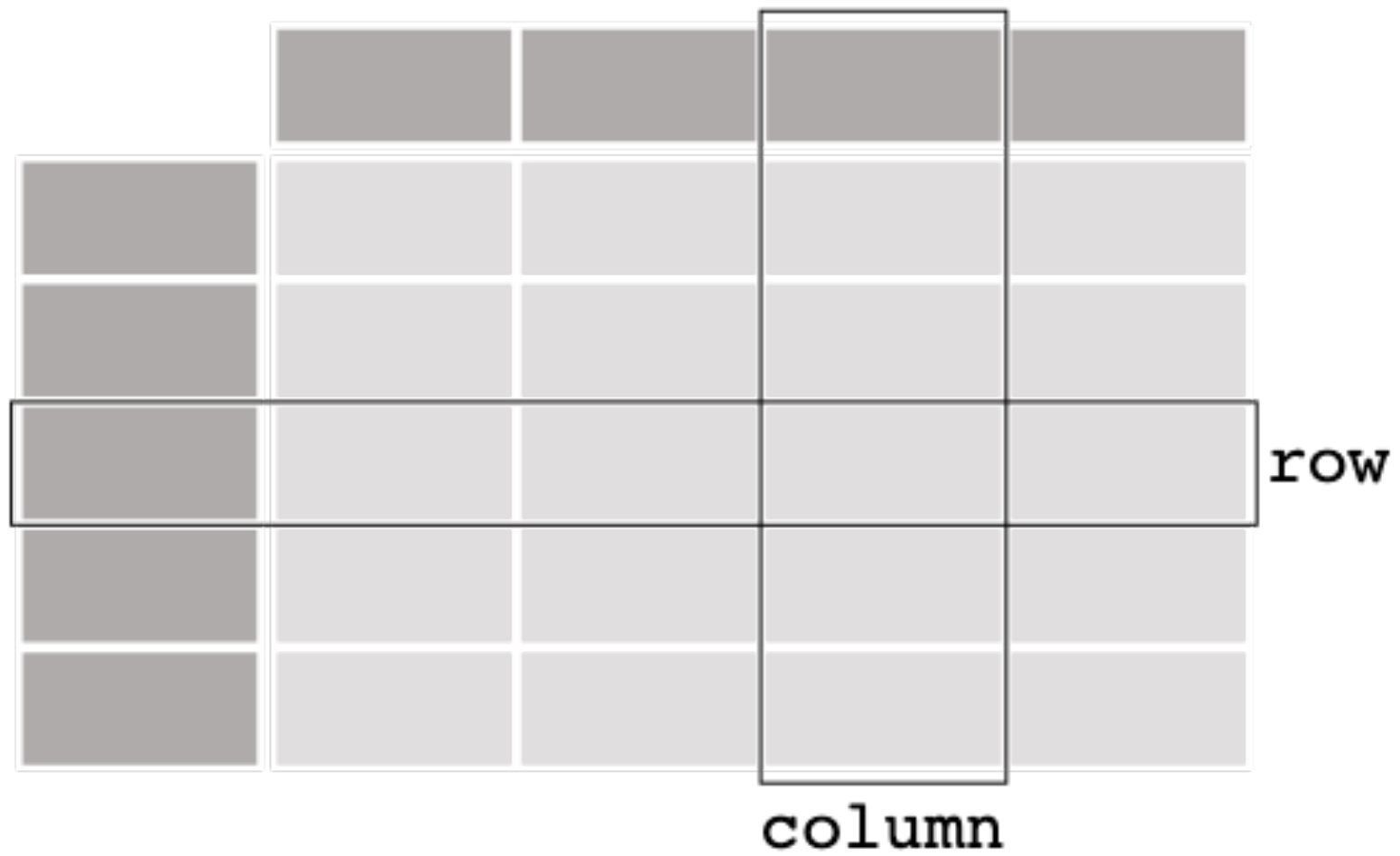
# Pandas

## Pandas Data structure



# Pandas

## DataFrame



# Pandas

```
In [1]: # Import the pandas package
import pandas as pd
```

```
In [2]: # Read airbnb dataset
airbnb = pd.read_csv('airbnb.csv')
```

```
In [3]: # Print pandas DataFrame
airbnb
```

Out[3]:

	listing_id	5_stars	availability_365	borough	coordinates	host_id	host_name	is Rated	last_review	name	neighbourhood	number_of_reviews
0	3831	0.757366	194	Brooklyn	(40.68514, -73.95976)	4869	LisaRoxanne	1	2019-07-05	Cozy Entire Floor of Brownstone	Clinton Hill	10
1	6848	0.789743	46	Brooklyn	(40.70837, -73.95352)	15991	Allen & Irina	1	2019-06-29	Only 2 stops to Manhattan studio	Williamsburg	10
2	7322	0.669873	12	Manhattan	(40.74192, -73.99501)	18946	Doti	1	2019-07-01	Chelsea Perfect	Chelsea	10
3	7726	0.640251	21	Brooklyn	(40.67592, -73.94694)	20950	Adam And Charity	1	2019-06-22	Hip Historic Brownstone Apartment with Backyard	Crown Heights	10
4	12303	0.918593	311	Brooklyn	(40.69673, -73.97584)	47618	Yolande	1	2018-09-30	1bdr w private bath. in lofty apt	Fort Greene	10
...	...	...	...	...	...	...	...	...	...	...	...	...

# Pandas

**Columns**

**Rows**

	<i>Name</i>	<i>Team</i>	<i>Number</i>	<i>Position</i>	<i>Age</i>
0	Avery Bradley	Boston Celtics	0.0	PG	25.0
1	John Holland	Boston Celtics	30.0	SG	27.0
2	Jonas Jerebko	Boston Celtics	8.0	PF	29.0
3	Jordan Mickey	Boston Celtics	NaN	PF	21.0
4	Terry Rozier	Boston Celtics	12.0	PG	22.0
5	Jared Sullinger	Boston Celtics	7.0	C	NaN
6	Evan Turner	Boston Celtics	11.0	SG	27.0

**Data**

GG



# Pandas

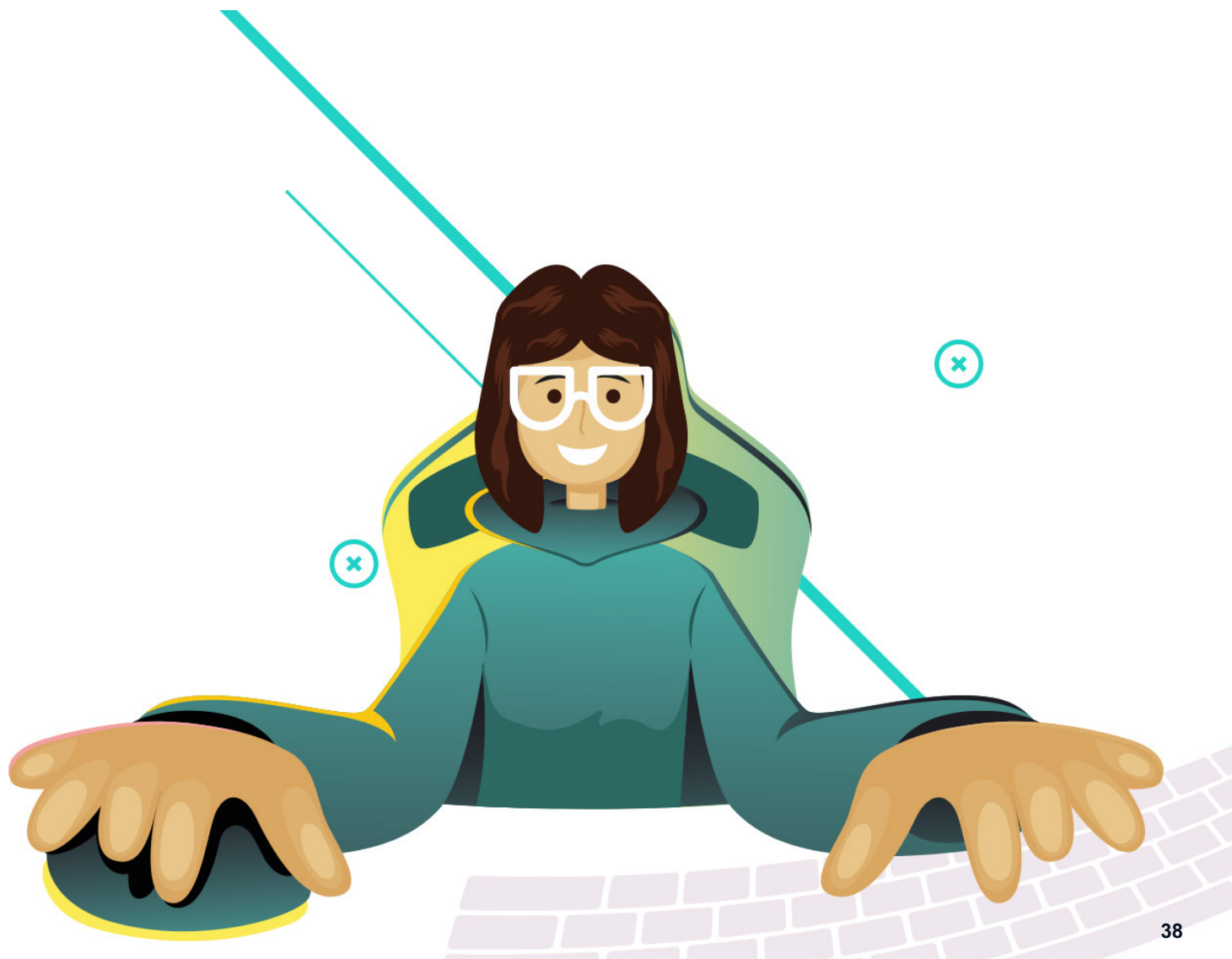
## ¿Qué podemos hacer?

- **Seleccionar ficheros**
- **Dibujar gráficos**
- **Combinar columnas**
- **Preparar y limpiar datos**
- ...

# Ejercicios



## Notebook 2

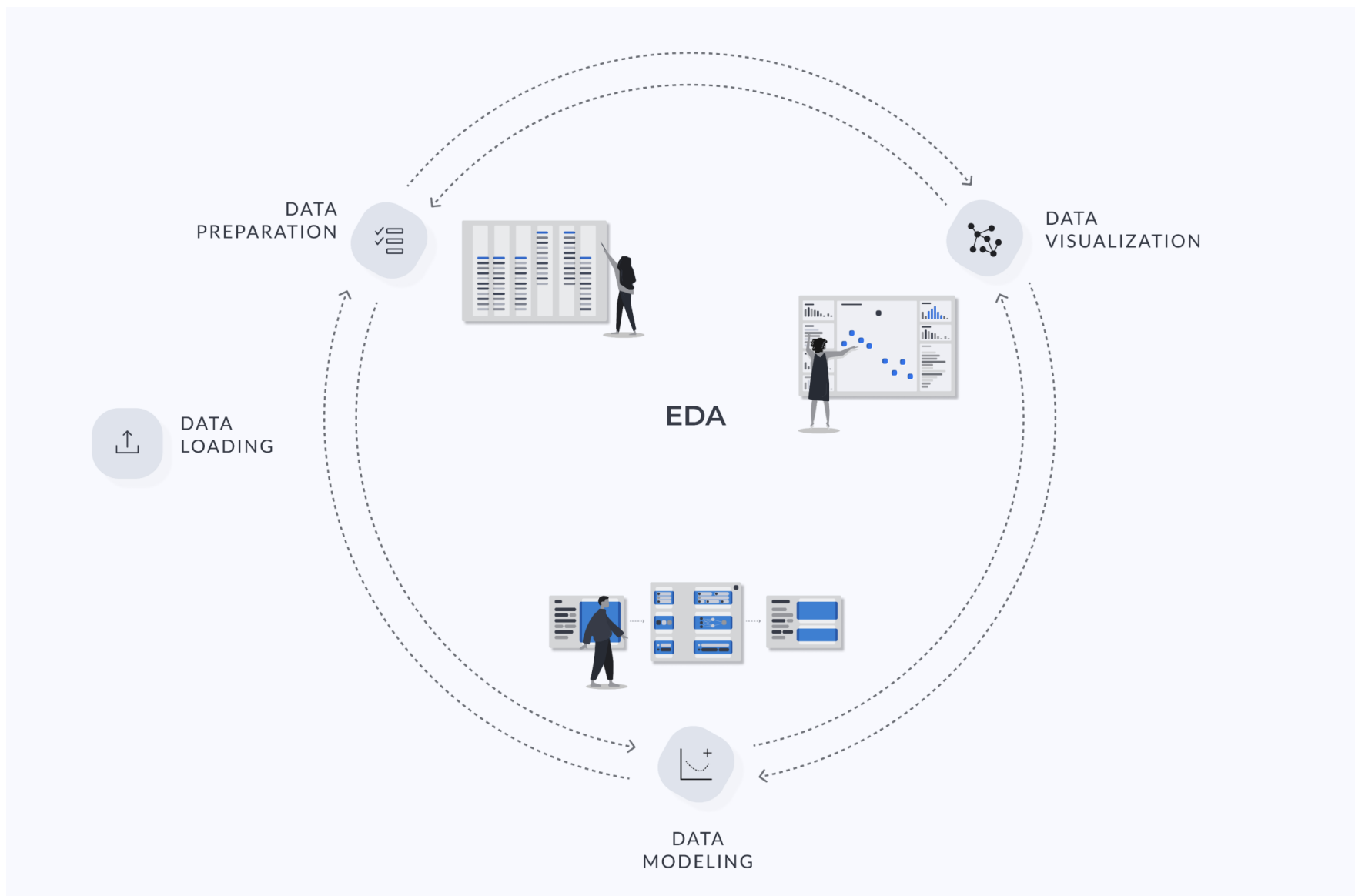


Manipulación de datos en Pandas

# Visualización



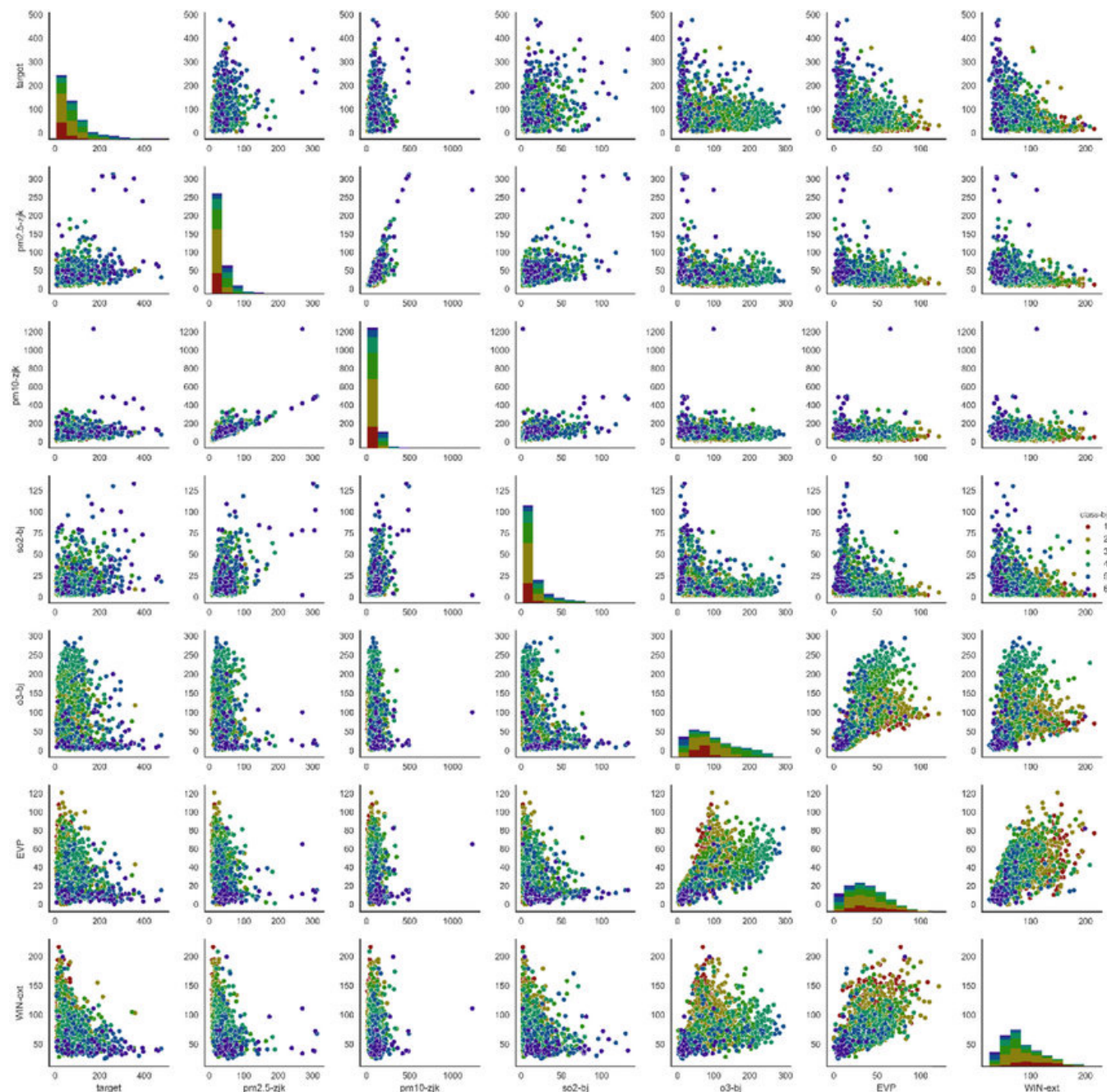
# Visualización



# Visualización



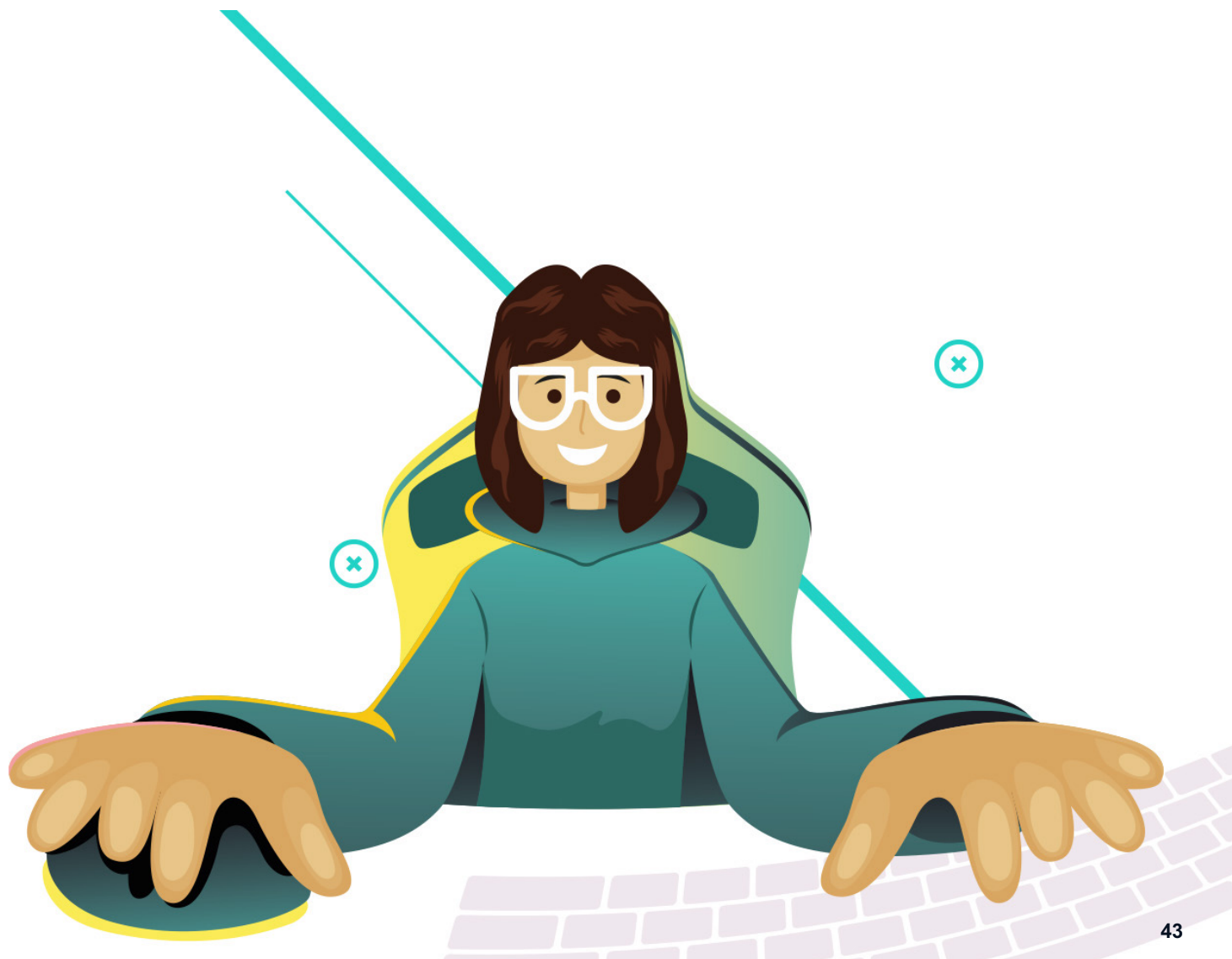
# Visualización



# Ejercicios



## Notebook 3





## Contacto

Correo: [a.cobo.aguilera@gmail.com](mailto:a.cobo.aguilera@gmail.com)

LinkedIn: [Aurora Cobo Aguilera](#)

GitHub: [AuroraCoboAguilera](#)

Google Scholar: [Aurora Cobo Aguilera](#)





red.es

Centro de  
Referencia Nacional  
en Comercio Electrónico  
y Marketing

CRN  
Digital



UNIÓN EUROPEA

*"El FSE invierte en tu futuro"*

**Fondo Social Europeo**

