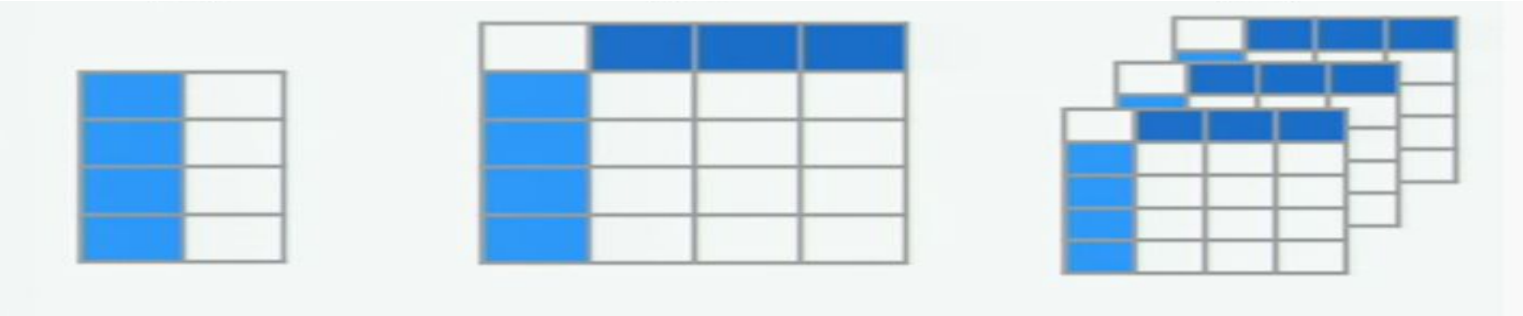
Introducción a Pandas

75.06 / 95.58 Organización de Datos, 2020-1C

¿Por qué Pandas?

- Pandas provee un entorno rápido, flexible y fácil de manejar para manipular y analizar datos.
- Maneja tres estructuras principales:

Series (1D) DataFrames (2D) Panels (3D)



¿Por qué Pandas?

- Estructuras de datos (series, dataframes)
- Manejo de índices
- Opciones simples para lectura y escritura de datos
- Herramientas para filtrar, seleccionar y transformar los datos
- Integración con otras librerías para análisis de datos de python

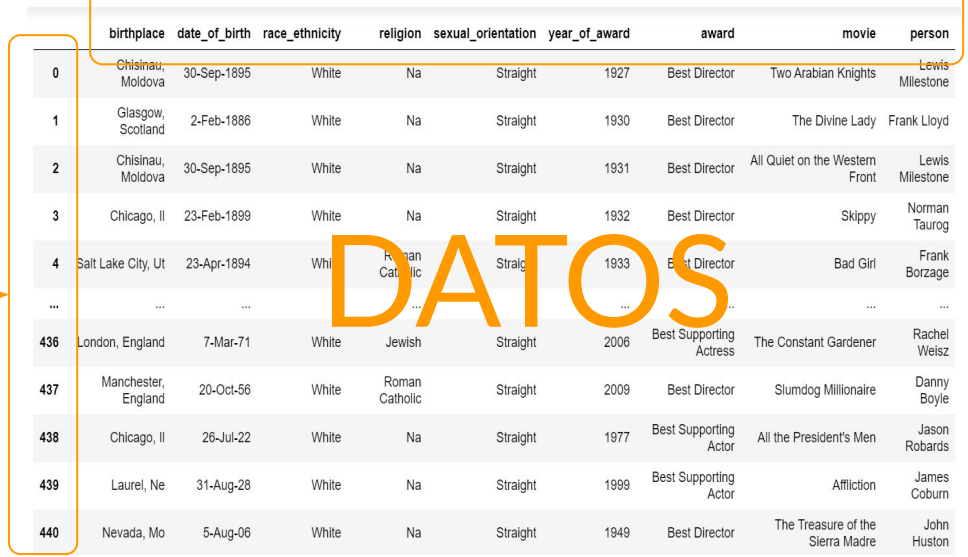
¿Por qué Pandas?

- Manejo de distintos tipos de datos
- Manejo simple de datos faltantes
- Fácil de codificar

DataFrames

Column Index

Index



	birthplace	date_of_birth	race_ethnicity	religion	sexual_orientation	year_of_award	award	movie	person
0	Chisinau, Moldova	30-Sep-1895	White	Na	Straight	1927	Best Director	Two Arabian Knights	Lewis Milestone
1	Glasgow, Scotland	2-Feb-1886	White	Na	Straight	1930	Best Director	The Divine Lady	Frank Lloyd
2	Chisinau, Moldova	30-Sep-1895	White	Na	Straight	1931	Best Director	All Quiet on the Western Front	Lewis Milestone
3	Chicago, Il	23-Feb-1899	White	Na	Straight	1932	Best Director	Skippy	Norman Taurog
4	Salt Lake City, Ut	23-Apr-1894	Whi	Roman Catholic	Straig	1933	Best Director	Bad Girl	Frank Borzage
...	...	...	...	...	...	...	...	...	...
436	London, England	7-Mar-71	White	Jewish	Straight	2006	Best Supporting Actress	The Constant Gardener	Rachel Weisz
437	Manchester, England	20-Oct-56	White	Roman Catholic	Straight	2009	Best Director	Slumdog Millionaire	Danny Boyle
438	Chicago, Il	26-Jul-22	White	Na	Straight	1977	Best Supporting Actor	All the President's Men	Jason Robards
439	Laurel, Ne	31-Aug-28	White	Na	Straight	1999	Best Supporting Actor	Affliction	James Coburn
440	Nevada, Mo	5-Aug-06	White	Na	Straight	1949	Best Director	The Treasure of the Sierra Madre	John Huston

DATOS

DataFrames: Indexes

	birthplace	date_of_birth	race_ethnicity	religion	sexual_orientation	year_of_award	award	movie	person
0	Chisinau, Moldova	30-Sep-1895	White	Na	Straight	1927	Best Director	Two Arabian Knights	Lewis Milestone
1	Glasgow, Scotland	2-Feb-1886	White	Na	Straight	1930	Best Director	The Divine Lady	Frank Lloyd
2	Chisinau, Moldova	30-Sep-1895	White	Na	Straight	1931	Best Director	All Quiet on the Western Front	Lewis Milestone
3	Chicago, Il	23-Feb-1899	White	Na	Straight	1932	Best Director	Skippy	Norman Taurog
4	Salt Lake City, Ut	23-Apr-1894	White	Roman Catholic	Straight	1933	Best Director	Bad Girl	Frank Borzage
...	...	...	...	...	...	...	...	...	...
436	London, England	7-Mar-71	White	Jewish	Straight	2006	Best Supporting Actress	The Constant Gardener	Rachel Weisz
437	Manchester, England	20-Oct-56	White	Roman Catholic	Straight	2009	Best Director	Slumdog Millionaire	Danny Boyle
438	Chicago, Il	26-Jul-22	White	Na	Straight	1977	Best Supporting Actor	All the President's Men	Jason Robards
439	Laurel, Ne	31-Aug-28	White	Na	Straight	1999	Best Supporting Actor	Affliction	James Coburn
440	Nevada, Mo	5-Aug-06	White	Na	Straight	1949	Best Director	The Treasure of the Sierra Madre	John Huston

DataFrames: Bloque de Datos

	birthplace	date_of_birth	race_ethnicity	religion	sexual_orientation	year_of_award	award	movie	person
0	Chisinau, Moldova	30-Sep-1895	White	Na	Straight	1927	Best Director	Two Arabian Knights	Lewis Milestone
1	Glasgow, Scotland	2-Feb-1886	White	Na	Straight	1930	Best Director	The Divine Lady	Frank Lloyd
2	Chisinau, Moldova	30-Sep-1895	White	Na	Straight	1931	Best Director	All Quiet on the Western Front	Lewis Milestone
3	Chicago, Il	23-Feb-1899	White	Na	Straight	1932	Best Director	Skippy	Norman Taurog
4	Salt Lake City, Ut	23-Apr-1894	White	Roman Catholic	Straight	1933	Best Director	Bad Girl	Frank Borzage
...	...	...	...	...	...	...	...	...	...
436	London, England	7-Mar-71	White	Jewish	Straight	2006	Best Supporting Actress	The Constant Gardener	Rachel Weisz
437	Manchester, England	20-Oct-56	White	Roman Catholic	Straight	2009	Best Director	Slumdog Millionaire	Danny Boyle
438	Chicago, Il	26-Jul-22	White	Na	Straight	1977	Best Supporting Actor	All the President's Men	Jason Robards
439	Laurel, Ne	31-Aug-28	White	Na	Straight	1999	Best Supporting Actor	Affliction	James Coburn
440	Nevada, Mo	5-Aug-06	White	Na	Straight	1949	Best Director	The Treasure of the Sierra Madre	John Huston

DataFrames: Columnas (Series)

	birthplace	late_of_birth	race_ethnicity	religion	sexual_orientation	year_of_award	award	movie	person
0	Chisinau, Moldova	30-Sep-1895	White	Na	Straight	1927	Best Director	Two Arabian Knights	Lewis Milestone
1	Glasgow, Scotland	2-Feb-1886	White	Na	Straight	1930	Best Director	The Divine Lady	Frank Lloyd
2	Chisinau, Moldova	30-Sep-1895	White	Na	Straight	1931	Best Director	All Quiet on the Western Front	Lewis Milestone
3	Chicago, Il	23-Feb-1899	White	Na	Straight	1932	Best Director	Skippy	Norman Taurog
4	Salt Lake City, Ut	23-Apr-1894	White	Roman Catholic	Straight	1933	Best Director	Bad Girl	Frank Borzage
...	...	...	...	...	...	...	...	...	...
436	London, England	7-Mar-71	White	Jewish	Straight	2006	Best Supporting Actress	The Constant Gardener	Rachel Weisz
437	Manchester, England	20-Oct-56	White	Roman Catholic	Straight	2009	Best Director	Slumdog Millionaire	Danny Boyle
438	Chicago, Il	26-Jul-22	White	Na	Straight	1977	Best Supporting Actor	All the President's Men	Jason Robards
439	Laurel, Ne	31-Aug-28	White	Na	Straight	1999	Best Supporting Actor	Affliction	James Coburn
440	Nevada, Mo	5-Aug-06	White	Na	Straight	1949	Best Director	The Treasure of the Sierra Madre	John Huston

Indexes

Indexes: Acceso a Datos

	birthplace	date_of_birth	race_ethnicity	religion	sexual_orientation	year_of_award	award	movie	person
0	Chisinau, Moldova	30-Sep-1895	White	Na	Straight	1927	Best Director	Two Arabian Knights	Lewis Milestone
1	Glasgow, Scotland	2-Feb-1886	White	Na	Straight	1930	Best Director	The Divine Lady	Frank Lloyd
2	Chisinau, Moldova	30-Sep-1895	White	Na	Straight	1931	Best Director	All Quiet on the Western Front	Lewis Milestone
3	Chicago, Il	23-Feb-1899	White	Na	Straight	1932	Best Director	Skippy	Norman Taurog
4	Salt Lake City, Ut	23-Apr-1894	White	Roman Catholic	Straight	1933	Best Director	Bad Girl	Frank Borzage
...	...	...	...	...	...	...	...	...	...
436	London, England	7-Mar-71	White	Jewish	Straight	2006	Best Supporting Actress	The Constant Gardener	Rachel Weisz
437	Manchester, England	20-Oct-56	White	Roman Catholic	Straight	2009	Best Director	Slumdog Millionaire	Danny Boyle
438	Chicago, Il	26-Jul-22	White	Na	Straight	1977	Best Supporting Actor	All the President's Men	Jason Robards
439	Laurel, Ne	31-Aug-28	White	Na	Straight	1999	Best Supporting Actor	Affliction	James Coburn
440	Nevada, Mo	5-Aug-06	White	Na	Straight	1949	Best Director	The Treasure of the Sierra Madre	John Huston

```
oscars.loc[0]
```

```
birthplace                Chisinau, Moldova
date_of_birth                   30-Sep-1895
race_ethnicity                        White
religion                                 Na
sexual_orientation                 Straight
year_of_award                          1927
award                         Best Director
movie                 Two Arabian Knights
person                    Lewis Milestone
Name: 0, dtype: object
```

Indexes: Facilitan la combinación de datos

person	date_of_birth	birthplace
Lewis Milestone	1895-09-30	Chisinau, Moldova
Frank Lloyd	1886-02-02	Glasgow, Scotland
Norman Taurog	1899-02-23	Chicago, Il
Frank Borzage	1894-04-23	Salt Lake City, Ut
Frank Capra	1897-05-18	Bisacquino, Sicily, Italy
...	...	...
Mo'Nique	2067-12-11	Woodlawn, Md
Melissa Leo	2060-09-14	New York City
Octavia Spencer	1972-05-25	Montgomery, Al
Anne Hathaway	1982-11-12	Brooklyn, Ny

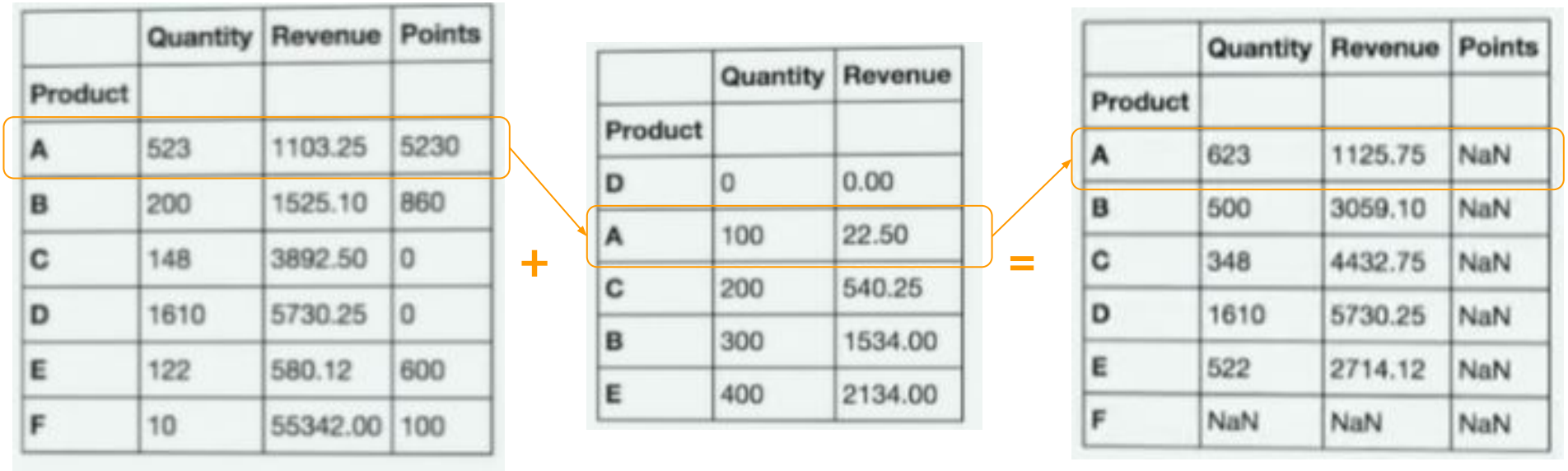
person	year_of_award	award	movie
Lewis Milestone	1927	Best Director	Two Arabian Knights
Frank Lloyd	1930	Best Director	The Divine Lady
Norman Taurog	1932	Best Director	Skippy
Frank Borzage	1933	Best Director	Bad Girl
Frank Capra	1935	Best Director	It Happened One Night
...	...	...	...
Mo'Nique	2010	Best Supporting Actress	Precious
Melissa Leo	2011	Best Supporting Actress	The Fighter
Octavia Spencer	2012	Best Supporting Actress	The Help
Anne Hathaway	2013	Best Supporting Actress	Les Misİ©rables
Lupita Nyong'o	2014	Best Supporting Actress	12 Years a Slave

Indexes: Facilitan la combinación de datos

	Quantity	Revenue	Points
Product			
A	523	1103.25	5230
B	200	1525.10	860
C	148	3892.50	0
D	1610	5730.25	0
E	122	580.12	600
F	10	55342.00	100

	Quantity	Revenue
Product		
D	0	0.00
A	100	22.50
C	200	540.25
B	300	1534.00
E	400	2134.00

Indexes: Facilitan la combinación de datos



Product	Quantity	Revenue	Points
A	523	1103.25	5230
B	200	1525.10	860
C	148	3892.50	0
D	1610	5730.25	0
E	122	580.12	600
F	10	55342.00	100

+

Product	Quantity	Revenue
D	0	0.00
A	100	22.50
C	200	540.25
B	300	1534.00
E	400	2134.00

=

Product	Quantity	Revenue	Points
A	623	1125.75	NaN
B	500	3059.10	NaN
C	348	4432.75	NaN
D	1610	5730.25	NaN
E	522	2714.12	NaN
F	NaN	NaN	NaN

Tipos de Datos

- Lectura de distintos tipos de archivos, y parseo de datos
  - Optimización de la lectura de CSVs
  - Librerías para lectura de otros formatos
- Conversión de tipos durante la lectura

```
unit_id,golden,unit_state,trusted_judgments,last_judgment_at,birthplace,birthplace:confidence,date_of_birth
670454353,FALSE,finalized,3,2/10/15 3:45,"Chisinau, Moldova",1,30-Sep-1895,1,White,1,Na,1,Straight,1,1927,1
670454354,FALSE,finalized,3,2/10/15 2:03,"Glasgow, Scotland",1,2-Feb-1886,1,White,1,Na,1,Straight,0.6842,19
670454355,FALSE,finalized,3,2/10/15 2:05,"Chisinau, Moldova",1,30-Sep-1895,1,White,1,Na,1,Straight,1,1931,0
670454356,FALSE,finalized,3,2/10/15 2:04,"Chicago, Il",1,23-Feb-1899,1,White,1,Na,1,Straight,1,1932,1,Best
```

Manejo de Bloques

Product ID	Description	Creation Date	Qty	Price	Provider	Last Update
A23	Pencil	21/5/2017	23	1,23	Anne	10/4/2020
B12	Eraser	13/8/2019	45	2,3	John	12/3/2020
A38	Red Pen	25/6/2018	31	3,45	Anne	19/3/2020
C27	Ruler	2/11/2018	67	2,19	Peter	7/4/2020
B33	Notebook	19/10/2019	29	4,99	Paul	16/3/2020

Product ID	Description	Creation Date	Qty	Price	Provider	Last Update
A23	Pencil	21/5/2017	23	1,23	Anne	10/4/2020
B12	Eraser	13/8/2019	45	2,3	John	12/3/2020
A38	Red Pen	25/6/2018	31	3,45	Anne	19/3/2020
C27	Ruler	2/11/2018	67	2,19	Peter	7/4/2020
B33	Notebook	19/10/2019	29	4,99	Paul	16/3/2020

Block Manager

Product ID
A23
B12
A38
C27
B33

Product ID
Description
Creation Date
Qty
Price
Provider
Last Update

Data Blocks

Object Block

	0	1
0	Pencil	Anne
1	Eraser	John
2	Red Pen	Anne
3	Ruler	Peter
4	Notebook	Paul

Dates Block

	0	1
0	21/5/2017	10/4/2020
1	13/8/2019	12/3/2020
2	25/6/2018	19/3/2020
3	2/11/2018	7/4/2020
4	19/10/2019	16/3/2020

Int Block

	0
0	23
1	45
2	31
3	67
4	29

Float Block

	0
0	1,23
1	2,3
2	3,45
3	2,19
4	4,99

Product ID	Description	Creation Date	Qty	Price	Provider	Last Update
A23	Pencil	21/5/2017	23	1,23	Anne	10/4/2020
B12	Eraser	13/8/2019	45	2,3	John	12/3/2020
A38	Red Pen	25/6/2018	31	3,45	Anne	19/3/2020
C27	Ruler	2/11/2018	67	2,19	Peter	7/4/2020
B33	Notebook	19/10/2019	29	4,99	Paul	16/3/2020

Block Manager

Product ID
A23
B12
A38
C27
B33

Product ID
Description
Creation Date
Qty
Price
Provider
Last Update

```
df.loc['A23',['Description','Provider']]
```

Data Blocks

Object Block

	0	1
0	Pencil	Anne
1	Eraser	John
2	Red Pen	Anne
3	Ruler	Peter
4	Notebook	Paul

Dates Block

	0	1
0	21/5/2017	10/4/2020
1	13/8/2019	12/3/2020
2	25/6/2018	19/3/2020
3	2/11/2018	7/4/2020
4	19/10/2019	16/3/2020

Int Block

	0
0	23
1	45
2	31
3	67
4	29

Float Block

	0
0	1,23
1	2,3
2	3,45
3	2,19
4	4,99

Product ID	Description	Creation Date	Qty	Price	Provider	Last Update
A23	Pencil	21/5/2017	23	1,23	Anne	10/4/2020
B12	Eraser	13/8/2019	45	2,3	John	12/3/2020
A38	Red Pen	25/6/2018	31	3,45	Anne	19/3/2020
C27	Ruler	2/11/2018	67	2,19	Peter	7/4/2020
B33	Notebook	19/10/2019	29	4,99	Paul	16/3/2020

Block Manager

Product ID
A23
B12
A38
C27
B33

Product ID
Description
Creation Date
Qty
Price
Provider
Last Update

```
df.loc['A23',['Description','Provider','Price']]
```

Data Blocks

Object Block

	0	1
0	Pencil	Anne
1	Eraser	John
2	Red Pen	Anne
3	Ruler	Peter
4	Notebook	Paul

Dates Block

	0	1
0	21/5/2017	10/4/2020
1	13/8/2019	12/3/2020
2	25/6/2018	19/3/2020
3	2/11/2018	7/4/2020
4	19/10/2019	16/3/2020

Int Block

	0
0	23
1	45
2	31
3	67
4	29

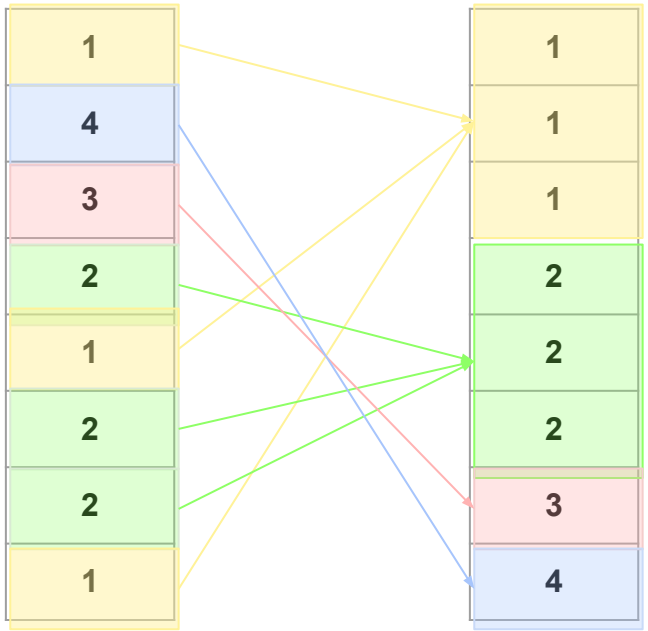
Float Block

	0
0	1,23
1	2,3
2	3,45
3	2,19
4	4,99

Factorización y Group By

- Factorización -> Mapeo de keys a ints
- Algoritmos mas simples y eficientes
  - No deben tener en cuenta data types.
- Utilizado para Groupby, hierarchical indexes, y datos categóricos.

Counting Sort



- Busca evitar la comparación de todos los elementos entre sí
- Obtener una lista de elementos únicos y sus cantidades
- Permite calcular bins para cada elemento, su tamaño y su orden, sin realizar las comparaciones entre todos los elementos.