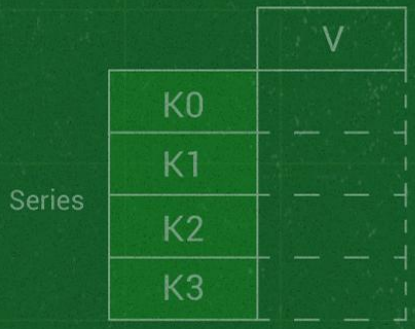
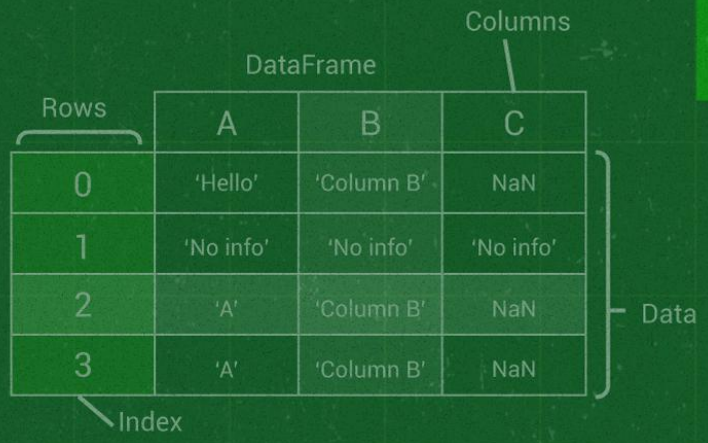
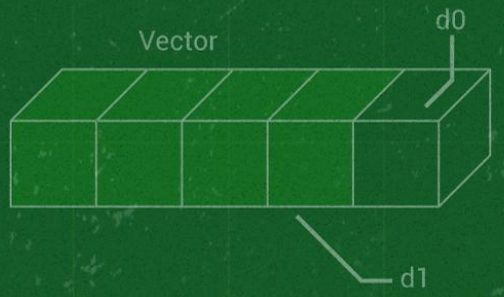
Curso Básico de

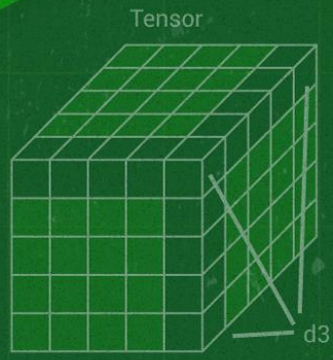
# Manipulación y Transformación de Datos con Pandas y NumPy

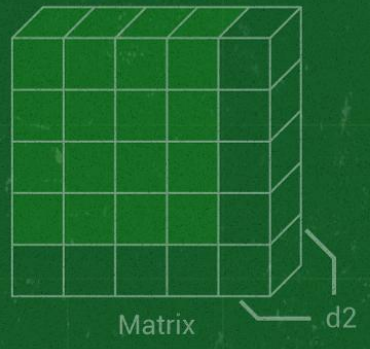
Carlos Alarcón



# Requisitos previos

- Fundamentos de matemáticas
- Python
- Jupyter Notebooks



# ¿Por qué NumPy y Pandas?

# ¿Para qué sirven?


- Cálculo numérico
- Manejo de arrays


- Manipulación y análisis de datos

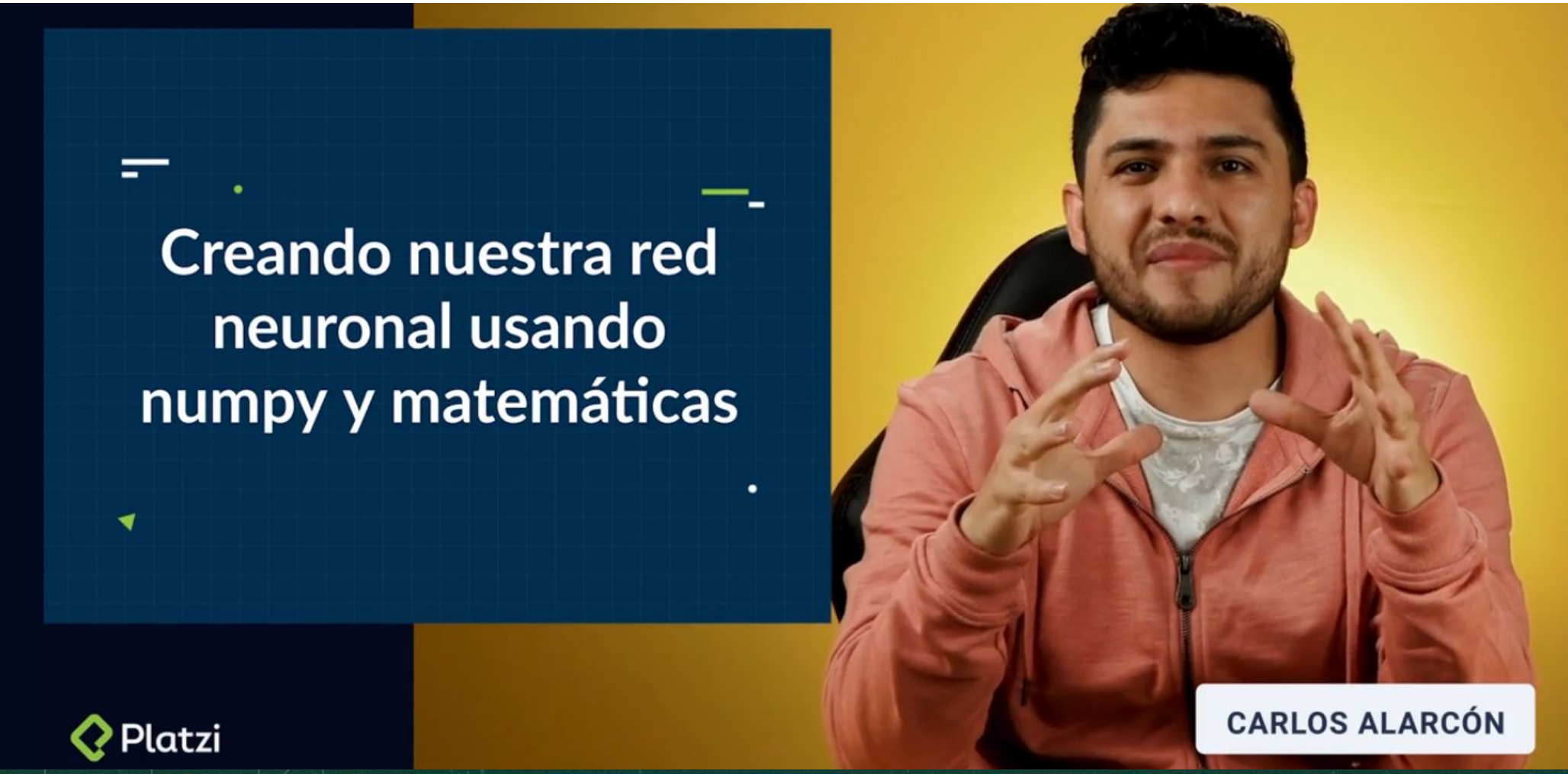
# NumPy

- Creada en el 2005.
- 100% open source.
- Se encuentra en GitHub.
- Travis Oliphant.

# ¿Por qué NumPy?

- Velocidad
- Tamaño
- Tipos de datos

# ¡Muy poderosa!



# Pandas

- Creada en el 2008.
- 100% open source.
- Se encuentra en GitHub.
- Wes McKinney.
- Construida sobre NumPy.

# ¿Por qué Pandas?

- Velocidad
- Poco código
- Múltiples formatos de archivos
- Alineación inteligente

# ¡Muy poderosa!

```
##Importando librerías

import numpy as np
import pandas as pd
```

# Los proyectos más grandes inician así

# NumPy Array

# Array

- La estructura central de NumPy.
- Representa datos de una manera estructurada.
- Indexado.
- Acceso a uno o muchos elementos.

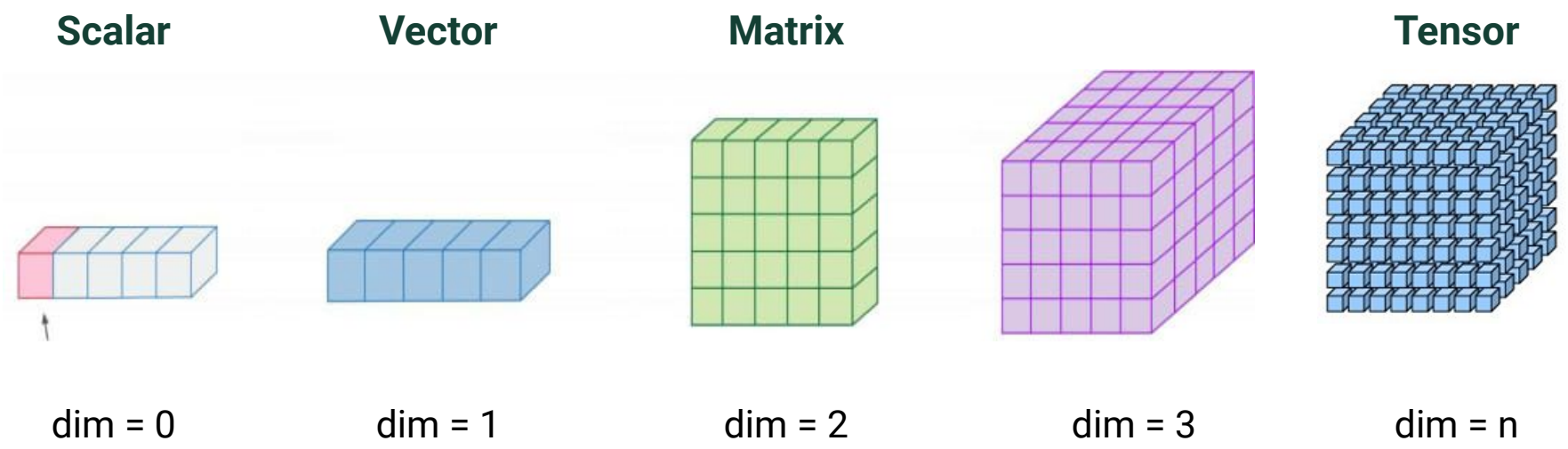
```
import numpy as np

a = np.array([1, 2, 3])
a
array([1, 2, 3])
```

# Tipos de datos

# Dimensiones

# Dimensiones de los datos



# Matrix - 2D

*(Ejemplos, características)*

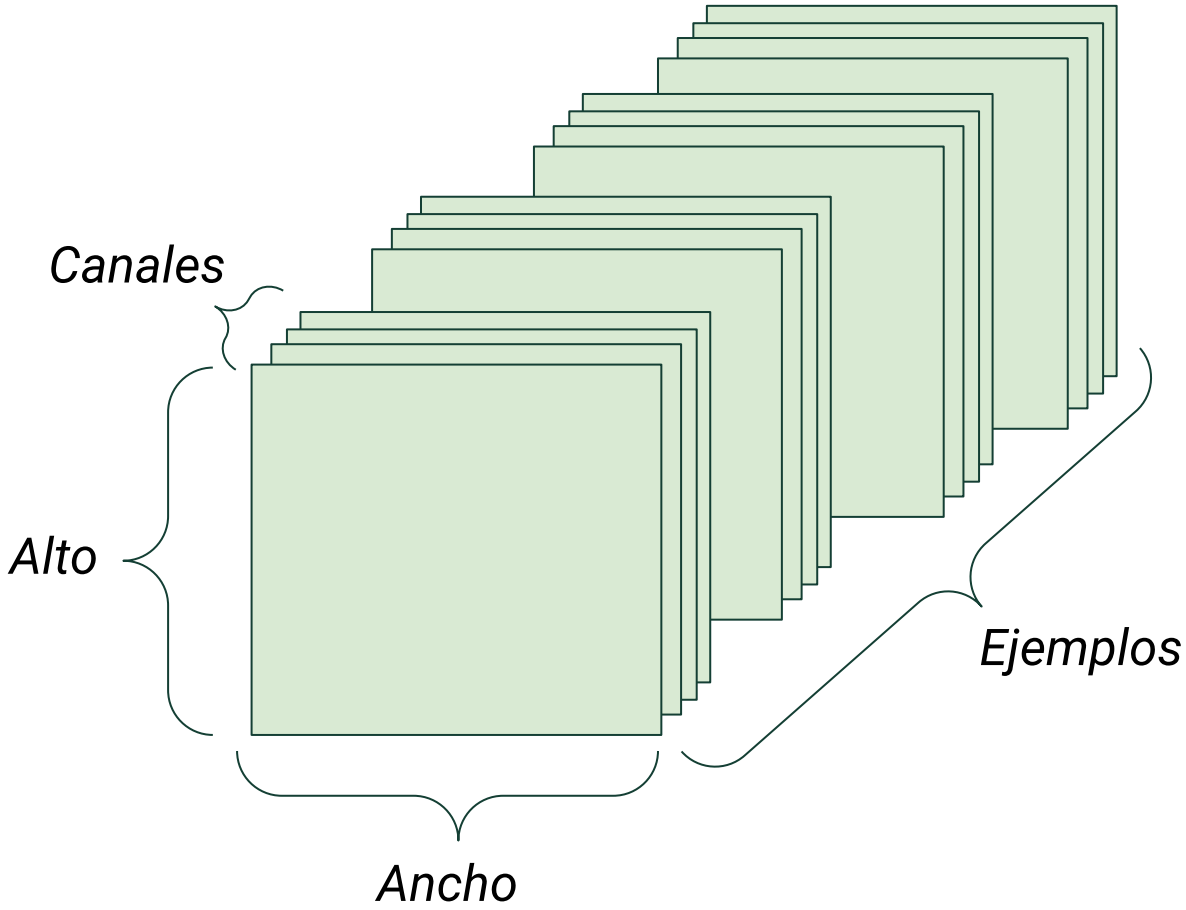
| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | **Order ID** | **Product** | **Category** | **Amount** | **Date** | **Country** |
| 2 | 1 | Carrots | Vegetables | $4,270 | 1/6/2012 | United States |
| 3 | 2 | Broccoli | Vegetables | $8,239 | 1/7/2012 | United Kingdom |
| 4 | 3 | Banana | Fruit | $617 | 1/8/2012 | United States |
| 5 | 4 | Banana | Fruit | $8,384 | 1/10/2012 | Canada |
| 6 | 5 | Beans | Vegetables | $2,626 | 1/10/2012 | Germany |
| 7 | 6 | Orange | Fruit | $3,610 | 1/11/2012 | United States |
| 8 | 7 | Broccoli | Vegetables | $9,062 | 1/11/2012 | Australia |
| 9 | 8 | Banana | Fruit | $6,906 | 1/16/2012 | New Zealand |
| 10 | 9 | Apple | Fruit | $2,417 | 1/16/2012 | France |
| 11 | 10 | Apple | Fruit | $7,431 | 1/16/2012 | Canada |
| 12 | 11 | Banana | Fruit | $8,250 | 1/16/2012 | Germany |
| 13 | 12 | Broccoli | Vegetables | $7,012 | 1/18/2012 | United States |
| 14 | 13 | Carrots | Vegetables | $1,903 | 1/20/2012 | Germany |

# Tensor - 3D

*(Ejemplos, Serie de tiempo, Características)*

# Tensor - 4D - Images

*(Ejemplos, Ancho, Alto, Canales de color)*



# Creando arrays

# Shape y Reshape

# Funciones principales de NumPy

# Copy

# Condiciones

# Operaciones

# Series y DataFrames

# Pandas Series

- Arreglo unidimensional indexado.
- Búsqueda por índice.
- Slicing.

| Índice | Datos |
|---|---|
| 1 | ‘A’ |
| 2 | ‘B’ |
| 3 | ‘C’ |
| 4 | ‘D’ |
| 5 | ‘E’ |

# Pandas Series

- Operaciones aritméticas.
- Distintos tipos de datos.

| Índice | Datos |
|---|---|
| 1 | ‘A’ |
| 2 | ‘B’ |
| 3 | ‘C’ |
| 4 | ‘D’ |
| 5 | ‘E’ |

# Pandas DataFrame

- Estructura principal.
- Arreglo de dos dimensiones.
- Búsqueda por índice (columnas o filas).

| | Key1 | Key2 | Key3 |
|---|---|---|---|
| **0** | 1 | Hank | 1.2 |
| **1** | 2 | Steve | 3.1 |
| **2** | 3 | Lisa | 3.1 |

# Pandas DataFrame

- Slicing
- Operaciones aritméticas
- Distintos tipos de datos
- Tamaño variable

| | Key1 | Key2 | Key3 |
|---|---|---|---|
| **0** | 1 | Hank | 1.2 |
| **1** | 2 | Steve | 3.1 |
| **2** | 3 | Lisa | 3.1 |

# Leer archivos con Pandas

# iloc y loc

# Agregar o eliminar datos con Pandas

# Manejo de datos nulos

# Filtrado por condiciones

# Funciones principales de Pandas

# Groupby

# Groupby

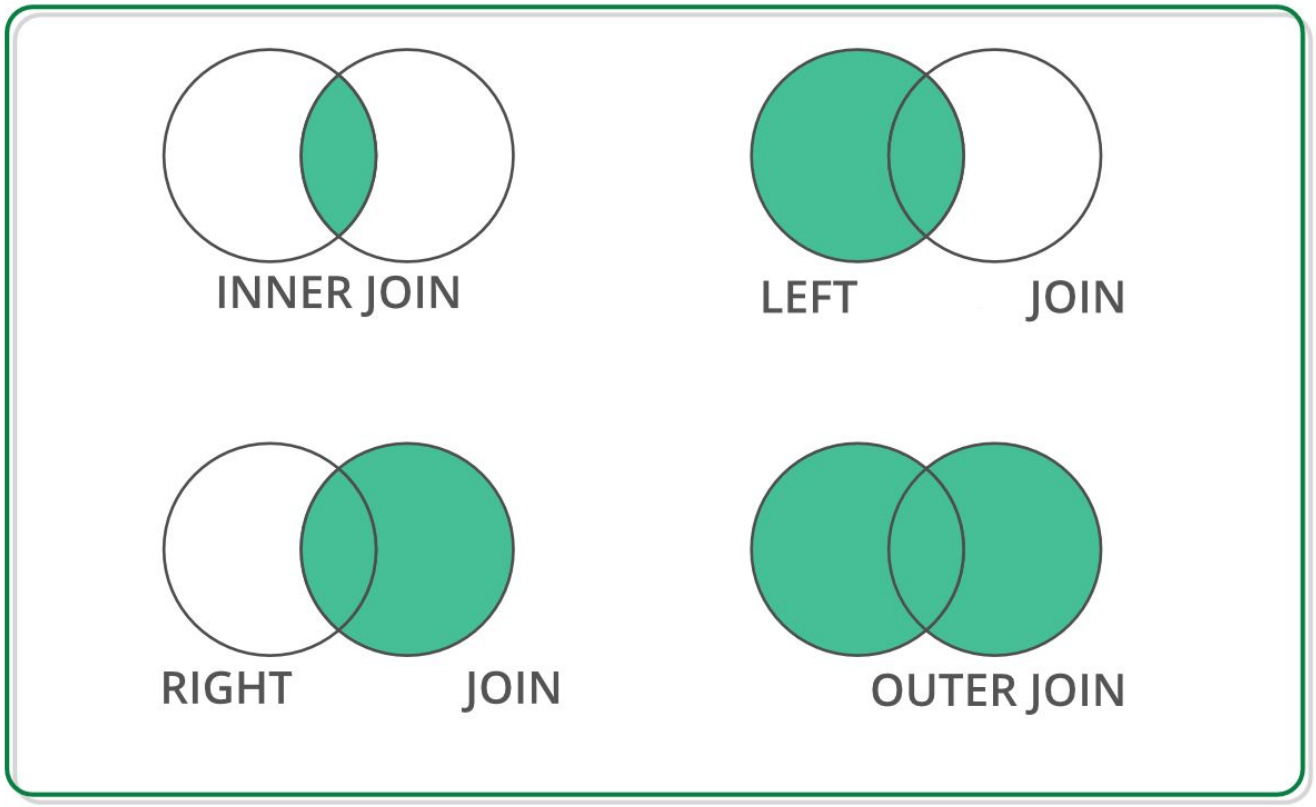
| Libro | Género | Precio |
|---|---|---|
| Libro 1 | Ciencia ficción | 10 |
| Libro 2 | Fantasía | 10 |
| Libro 3 | Fantasía | 10 |
| Libro 4 | Educación | 10 |
| Libro 5 | Ciencia ficción | 10 |
| Libro 6 | Ciencia ficción | 10 |

**Groupby Genero**

| Género | Sum | Count | Mean |
|---|---|---|---|
| Ciencia ficción | 30 | 3 | 10 |
| Fantasía | 20 | 2 | 10 |
| Educación | 10 | 1 | 10 |

# Combinando DataFrames

# Merge - Join



# Merge - Join

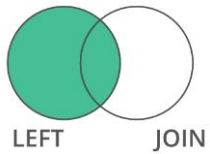
Left

| | A | B |
|---|---|---|
| 0 | | |
| 1 | | |

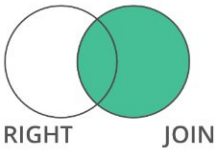
Merge

Right

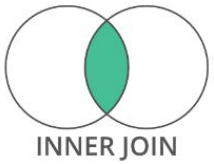
| | C | D |
|---|---|---|
| 0 | | |
| 2 | | |



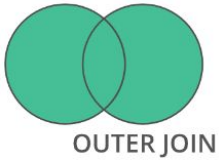

| | A | B | C | D |
|---|---|---|---|---|
| 0 | | | | |
| 1 | | | | |




| | A | B | C | D |
|---|---|---|---|---|
| 0 | | | | |
| 2 | | | | |




| | A | B | C | D |
|---|---|---|---|---|
| 0 | | | | |




| | A | B | C | D |
|---|---|---|---|---|
| 0 | | | | |
| 1 | | | | |
| 2 | | | | |

# Concat - Axis 0

**df1**

| | A | B | C | D |
|---|---|---|---|---|
| 0 | A0 | B0 | C0 | D0 |
| 1 | A1 | B1 | C1 | D1 |

**df2**

| | A | B | C | D |
|---|---|---|---|---|
| 2 | A2 | B2 | C2 | D2 |
| 3 | A3 | B3 | C3 | D3 |

**Concat**

| | A | B | C | D |
|---|---|---|---|---|
| 0 | A0 | B0 | C0 | D0 |
| 1 | A1 | B1 | C1 | D1 |
| 2 | A2 | B2 | C2 | D2 |
| 3 | A3 | B3 | C3 | D3 |

# Concat - Axis 1

**df1**

| | A | B | C | D |
|---|---|---|---|---|
| 0 | A0 | B0 | C0 | D0 |
| 1 | A1 | B1 | C1 | D1 |

**df2**

| | C | D | E | F |
|---|---|---|---|---|
| 1 | C1 | D1 | E1 | F1 |
| 2 | C2 | D2 | E2 | F2 |

**Concat**

| | A | B | C | D | C | D | E | F |
|---|---|---|---|---|---|---|---|---|
| 0 | A0 | B0 | C0 | D0 | NaN | NaN | NaN | NaN |
| 1 | A1 | B1 | C1 | D1 | C1 | D1 | E1 | F1 |
| 2 | NaN | NaN | NaN | NaN | C2 | D2 | E2 | F2 |

# Merge, Join y Concat

# Pivot, Melt

# Apply

# Posibilidades con Pandas y NumPy

# Numpy

- El Array (ventajas).
- Tipos de datos.
- Slicing.
- Index.
- Dimensiones.
- Crear arrays.
- Matemática en numpy.
- Filtrado.
- Shape - Reshape.

# Pandas

- Series - Dataframe
- Files
- iloc - loc
- Manejo de nulos (NaN)
- Funciones principales
- Merge, Concat, Join
- Apply

# Posibilidades infinitas

# Carlos Andrés Alarcón

@Alarcon7a