



DATA SCIENCE

UNIDAD 1

Introducción a Pandas

INTRODUCCIÓN A PANDAS





- Aprender las bases de pandas: para qué sirve, cuáles son sus estructuras de datos y aplicaciones principales.
- Aplicar los conocimientos adquiridos de pandas en prácticas guiadas y labs.

Repaso Numpy



- Fortalezas y limitaciones de **Numpy: debatir en clase**

Pandas



- Colección de estructuras de datos y funciones que facilitan el trabajo con datos estructurados.
- Nombre derivado de "Panel Data System" (término econométrico para datasets multidimensionales).

Pandas



- Construído en base a Numpy, inicialmente por Wes McKinney.
- Ofrece más versatilidad que Numpy: podemos trabajar con datos de diferentes tipos y podemos identificar las filas y columnas usando etiquetas, en lugar de solamente enteros.
- Combina la alta performance de las operaciones sobre arrays de NumPy con la flexibilidad en la manipulación de datos de un spreadsheet o una base de datos relacional.

Pandas



- Provee **funcionalidades de indexación avanzadas** para facilitar la manipulación, agregación y selección de partes de un dataset.
- Provee operaciones de agrupación por columnas, filtros y sumarizaciones.
- Estas características hacen de pandas una librería indispensable para las tareas de limpieza y preprocesamiento de datos.

Estructuras de datos de Pandas



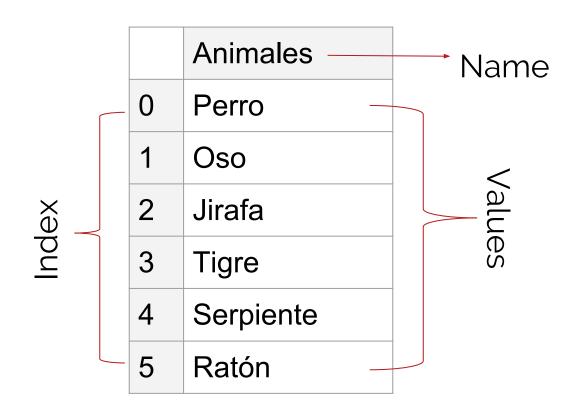
- Veremos los siguientes objetos:
 - Series
 - DataFrames
 - Index



	Animales	
0	Perro	
1	Oso	
2	Jirafa	
3	Tigre	
4	Serpiente	
5	Ratón	

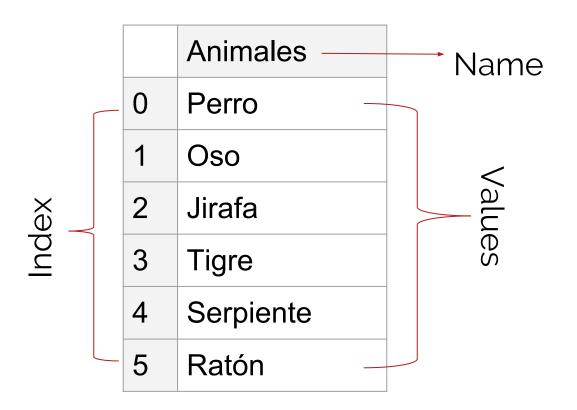
Una Series es un objeto similar a un **vector uni-dimensional**.





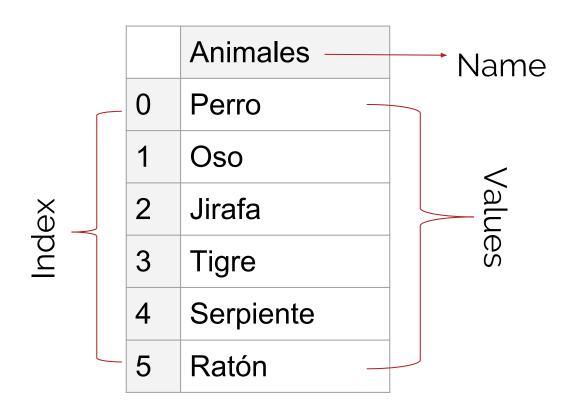
Contiene un array de valores y un array asociado de etiquetas de estos valores denominado como índice.





Una Serie también puede ser pensada como un diccionario de tamaño fijo con sus claves ordenadas.





Al igual que los arrays de NumPy, permiten pasar una **lista de valores con índices** para seleccionar un subconjunto de valores.

Pandas: DataFrame



13

Eje 1 (columnas)

Eje o (filas)

	Animales	Dueños
0	Perro	Juan
1	Oso	Pedro
2	Jirafa	Cristian
3	Tigre	Esteban
4	Serpiente	Pablo
5	Ratón	Claudio

Representa una estructura de datos tabular que contiene una **colección de** columnas, cada una de las cuales tiene un tipo determinado (number. string, boolean, etc.).

Pandas: DataFrame





El DataFrame tiene **índices de columnas y filas**.

CONCLUSIONES



Pandas: DataFrame



- En el día a día de un data scientist la limpieza, preparado y normalización de los datos con los que trabaja es la tarea que más tiempo insume.
- Pandas es la librería principal al momento de tener que realizar estas tareas.
- Los DataFrame de Pandas son la herramienta fundamental que vamos a utilizar a lo largo del curso y de nuestro día de trabajo a partir de ahora.