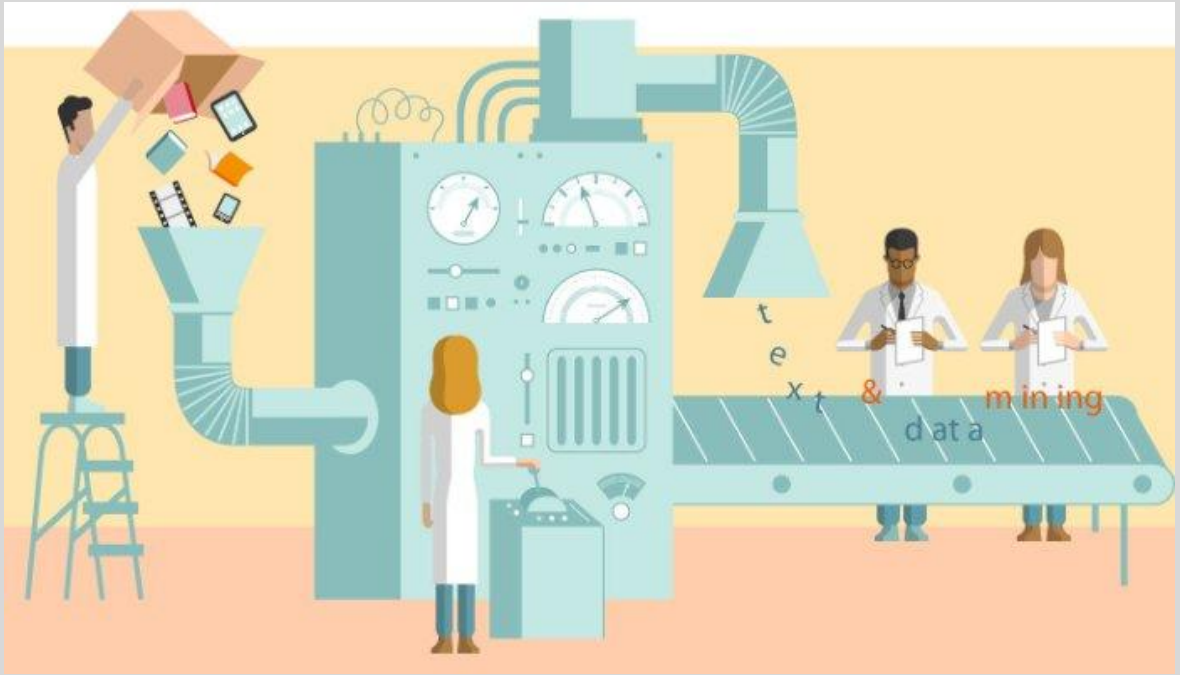


DigitalHouse > Coding School

DATA SCIENCE

UNIDAD 1
MÓDULO 2

Introducción a Pandas

Agosto 2017

INTRODUCCIÓN A PANDAS

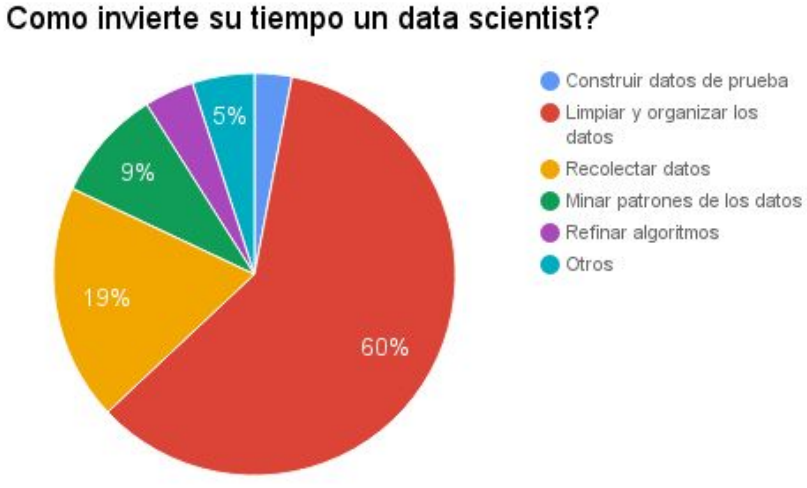
OBJETIVOS DE LA CLASE

1. Limpieza y preparación de datos
2. Qué es Pandas?
3. Para qué sirve Pandas?
4. Práctica guiada
5. Práctica independiente

Limpieza y preparación de datos

Limpieza y preparación de datos

- En 2009 Mike Driscoll (data scientist y CEO de Metamarkets) popularizó el término **“data munging”** para referirse al **arduo proceso de limpiar, preparar y validar los datos**



Fuente:
http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

Limpieza y preparación de datos

- **En 2013, Josh Wills** (ex director de Data Science de Cloudera y actual Director of Data Engineering en Slack) comenta: "**I'm a data janitor.** That's the sexiest job of the 21st century. It's very flattering, but it's also a little baffling."

Traducción:
En Data Science, se invierte un 80% del tiempo en preparar los datos y el 20% restante en quejarse de la necesidad de preparar los datos

Python vs otros para limpieza de datos

- Proceso reproducible
- Control de versiones
- Creación de tests automáticos
- Facilita el mantenimiento
- Lenguaje dinámico que permite alta productividad
- Posibilidad de interfacear con C para performance
- Preferido en ambientes de data science

http://lemire.me/blog/2014/05/23/you-shouldnt-use-a-spreadsheet-for-important-work-i-mean-it/

Pandas

Pandas

- **Colección de funciones y estructuras de datos** que facilitan el trabajo con datos estructurados
- Construído **en base a Numpy** inicialmente por Wes McKinney
- Nombre derivado de "**Pa**nel **Da**ta **S**ystem" (término econométrico para datasets multidimensionales)
- Brinda capacidades flexibles de manipulación de datos similares a spreadsheets y bases de datos relacionales

Pandas

- 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
- 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**

Objetos en Pandas

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

Pandas

- **En Pandas se utilizan dos estructuras de datos fundamentales *Series* y *DataFrames***
- Ambas estructuras usan arrays de **Numpy** como base
- Una *Series* es un array unidimensional capaz de guardar cualquier tipo de datos (enteros, strings, floats, objetos Python, etc.)
- Un *DataFrame* es una matriz bidimensional. Puede verse como un conjunto de Series que comparten todas los mismos valores en el índice.

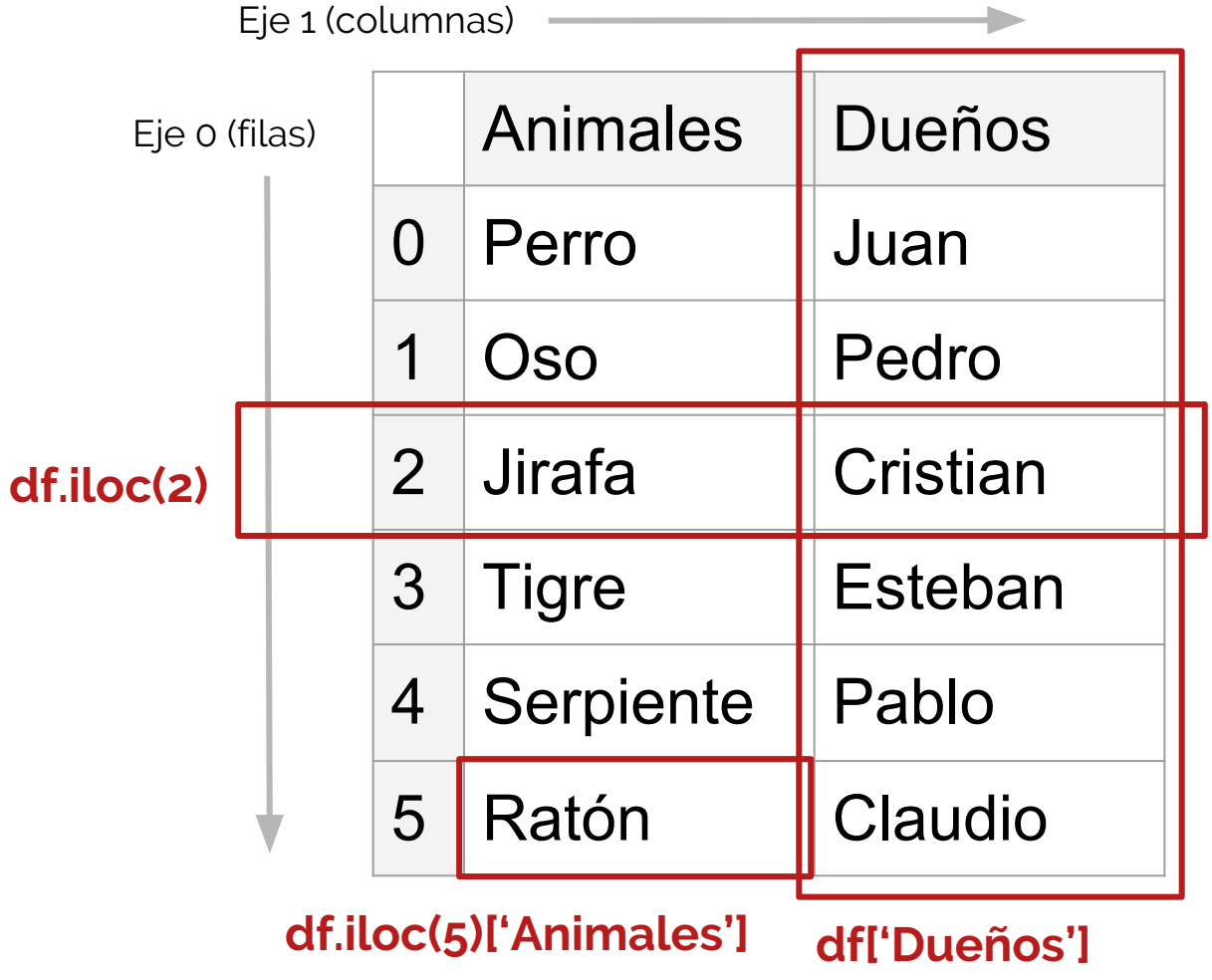
Pandas: Series

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

Pandas: Series

- 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 colección Series también puede ser considerado como un **diccionario de tamaño fijo con sus claves ordenadas**
- Comparados con los arrays de NumPy, permiten pasar una lista de valores del índice para seleccionar un subconjunto de valores

Pandas: DataFrame



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

Pandas: DataFrame

- 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.)
- Inspirados en el paquete data.frame de R
- Permiten operaciones “ricas” sobre índices como los JOIN y GROUP BY en SQL
- Ideales para organizar el resultado de un análisis en un formato útil para graficar el resultado o mostrarlo

Práctica Guiada 1
Objetos en Pandas

Indexación Booleana en Pandas

	Nombre	Apellido
0	Juan	Perez
1	Pedro	García
2	Matías	Zabala

True
False
True

	Nombre	Apellido
0	Juan	Perez
1	Matías	Zabala

Indexación y Subsetting en Pandas

- Para acceder y subsettear a los objetos en Pandas es necesario introducir los siguientes métodos:
  - .loc
  - .iloc

Práctica Guiada 2
Subsetting e Indexing de Objetos en Pandas

LAB: Pandas con datos de la Encuesta Permanente de Hogares

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
- Python es un lenguaje que numerosos features que nos pueden facilitar estas tareas
- 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