Bases de Datos

Clase 11: Pandas y ciencia de datos

Hasta ahora

- Bases de datos relacionales
- SQL

Ya tengo los datos guardados en una base y sé cómo hacer consultas… y ahora qué?

Data Science

Data science is a multidisciplinary approach to extracting actionable insights from the large and ever-increasing volumes of data collected and created by today’s organizations. Data science encompasses preparing data for analysis and processing, performing advanced data analysis, and presenting the results to reveal patterns and enable stakeholders to draw informed conclusions.

Según IBM

Data Science

Engloba varios conceptos:

- Data Analysis
- Machine Learning
- Deep Learning
- AI
- Computer Vision
- Natural Language Processing

y varios más …

Data Science

Engloba varios conceptos:

- **Data Analysis**

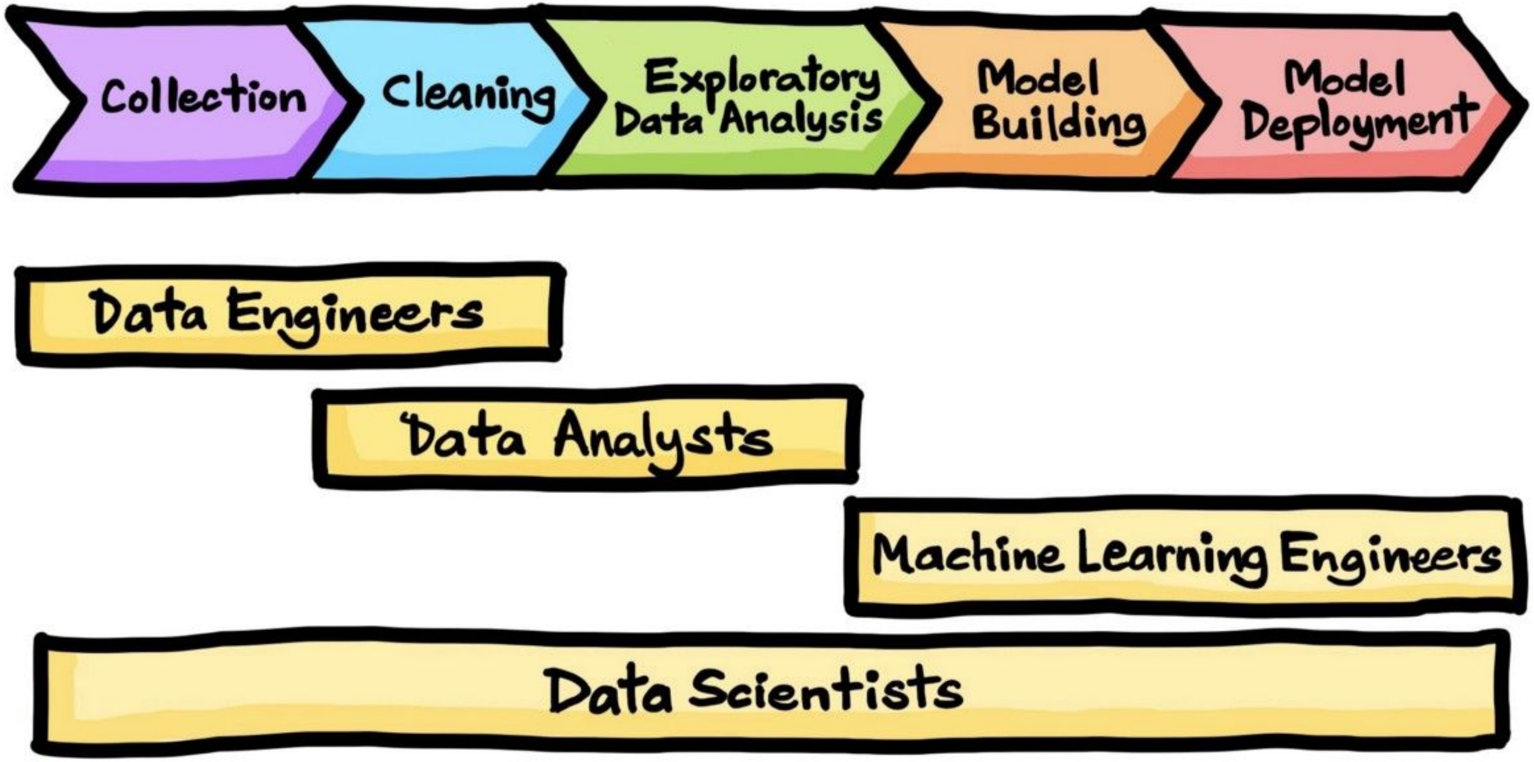
En esta clase veremos la librería más usada en Python para hacer Data Analysis: Pandas 🐼

En general, hay otros cursos que tocan los demás tópicos…

Data Analysis

Es el proceso de inspección, limpieza, transformación y modelamiento de datos que tiene como objetivo extraer información relevante para algún fin en particular.

El proceso de Data Science



En general, se ve así…

Tenemos los datos en una BD

Los cargamos en un entorno para trabajarlos (jupyter notebook, R, Tableau, etc)

Extraemos información importante con la cual se arman reportes, modelos, etc

La librería Pandas 🐼

pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.

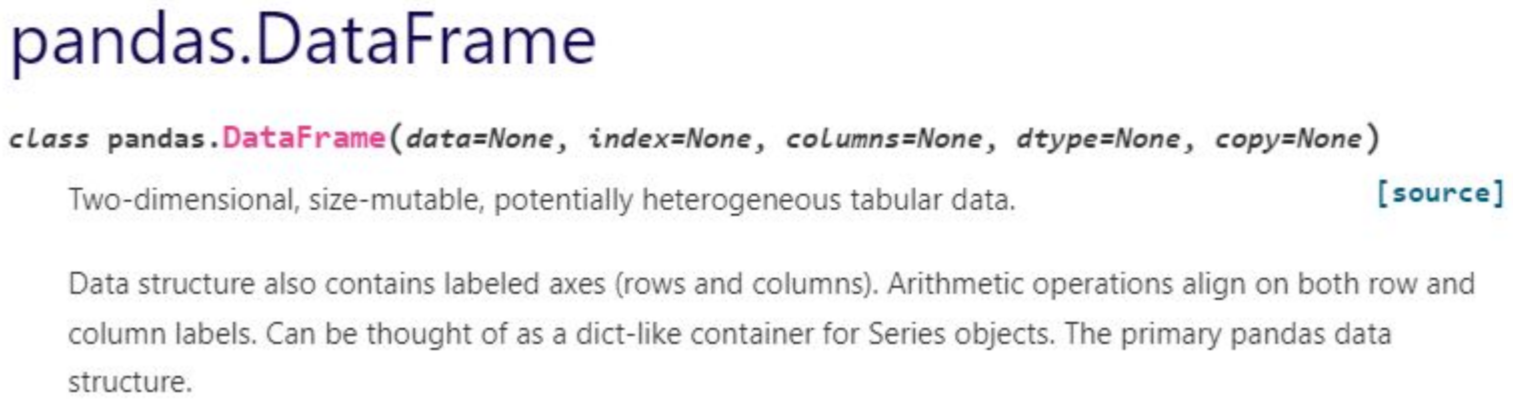
La librería Pandas 🐼

Soporta archivos en formato:

- .xlsx, .xls, …
- .csv
- .tsv
- .json

y otras más …
(https://pandas.pydata.org/docs/user_guide/io.html)

La librería Pandas: DataFrames



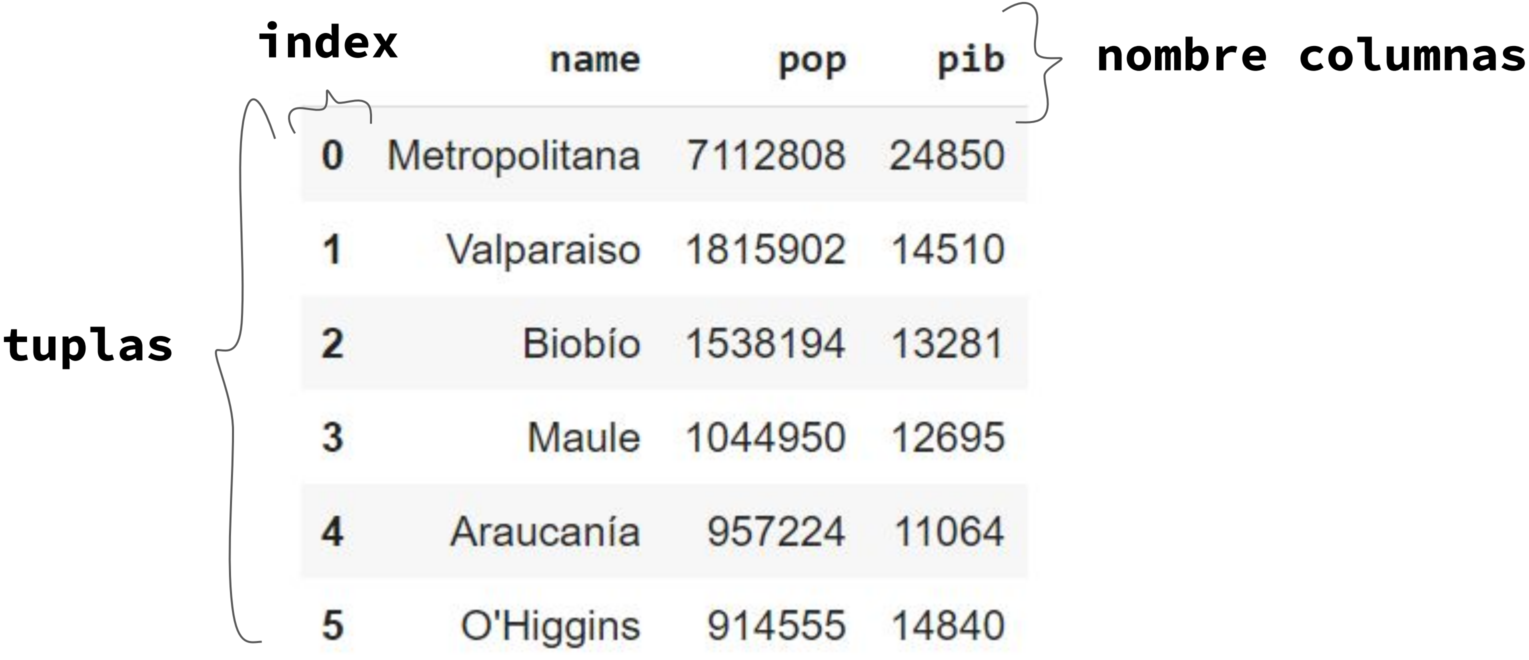
pandas.DataFrame

```
class pandas.DataFrame(data=None, index=None, columns=None, dtype=None, copy=None)
```

Two-dimensional, size-mutable, potentially heterogeneous tabular data. [source]

Data structure also contains labeled axes (rows and columns). Arithmetic operations align on both row and column labels. Can be thought of as a dict-like container for Series objects. The primary pandas data structure.

La librería Pandas: DataFrames



index

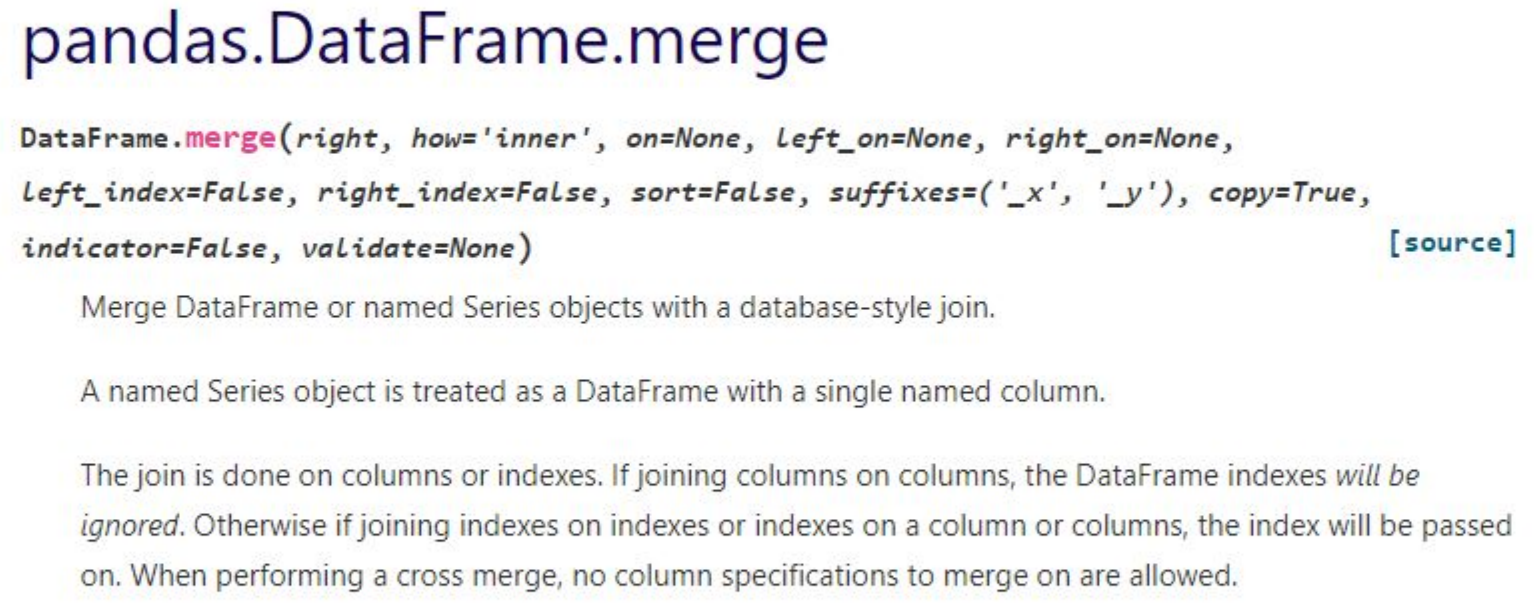
nombre columnas

tuplas

	name	pop	pib
0	Metropolitana	7112808	24850
1	Valparaiso	1815902	14510
2	Biobío	1538194	13281
3	Maule	1044950	12695
4	Araucanía	957224	11064
5	O'Higgins	914555	14840

La librería Pandas: DataFrames

Joins (Metodo Merge)



pandas.DataFrame.merge

```
DataFrame.merge(right, how='inner', on=None, left_on=None, right_on=None,
left_index=False, right_index=False, sort=False, suffixes=('_x', '_y'), copy=True,
indicator=False, validate=None)
```

[source]

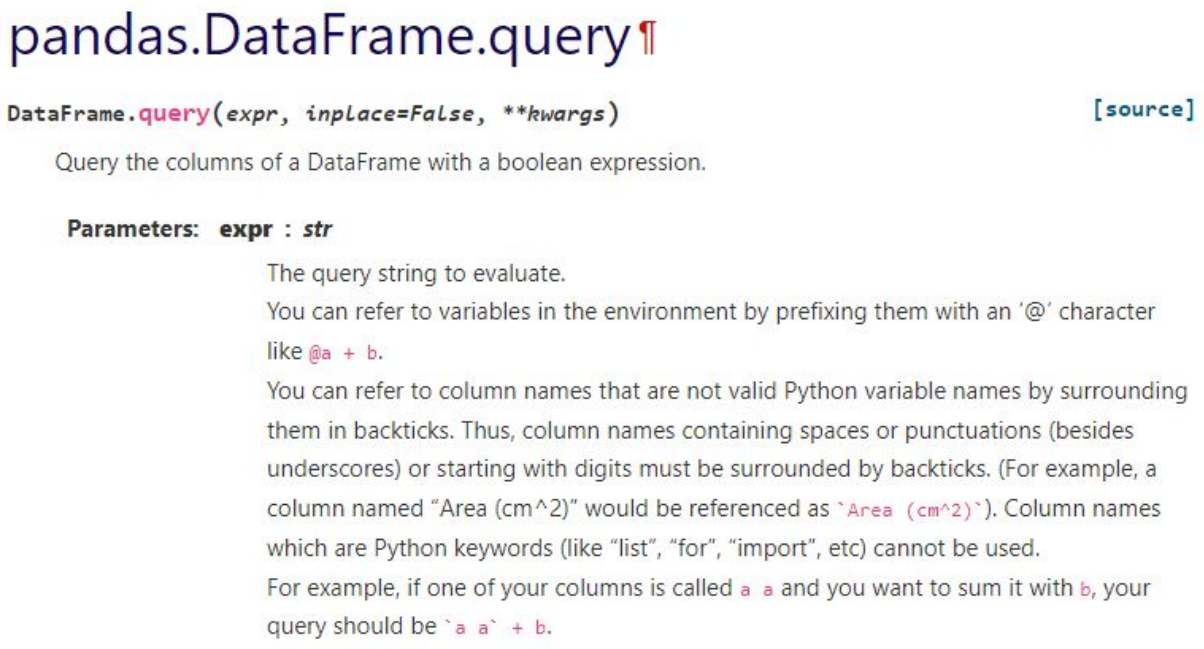
Merge DataFrame or named Series objects with a database-style join.

A named Series object is treated as a DataFrame with a single named column.

The join is done on columns or indexes. If joining columns on columns, the DataFrame indexes *will be ignored*. Otherwise if joining indexes on indexes or indexes on a column or columns, the index will be passed on. When performing a cross merge, no column specifications to merge on are allowed.

La librería Pandas: DataFrames

Query



pandas.DataFrame.query¶

`DataFrame.query(expr, inplace=False, **kwargs)` [source]

Query the columns of a DataFrame with a boolean expression.

Parameters: **expr** : *str*

The query string to evaluate.

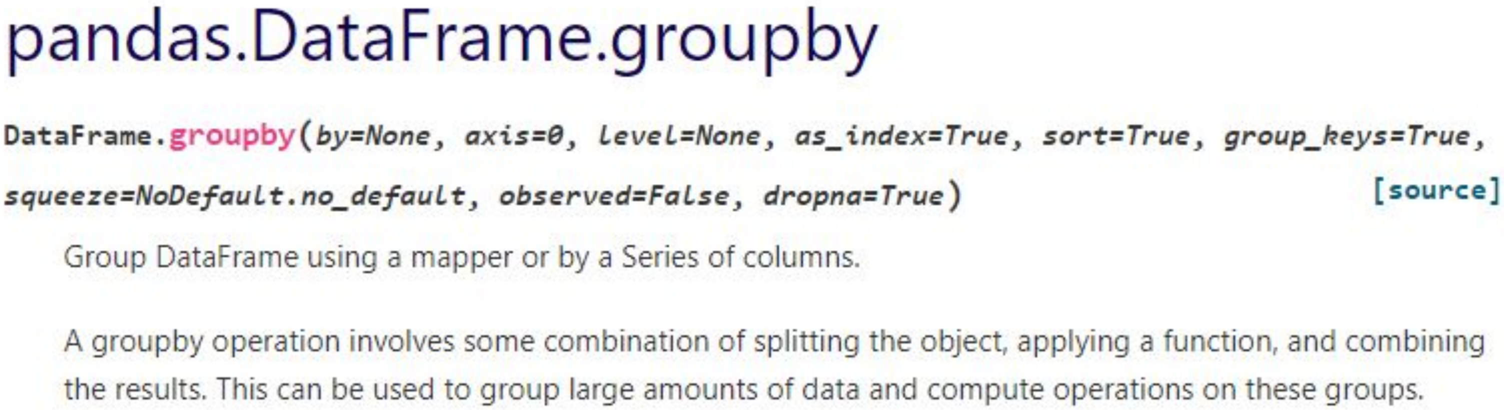
You can refer to variables in the environment by prefixing them with an '@' character like `@a + b`.

You can refer to column names that are not valid Python variable names by surrounding them in backticks. Thus, column names containing spaces or punctuations (besides underscores) or starting with digits must be surrounded by backticks. (For example, a column named "Area (cm^2)" would be referenced as `` `Area (cm^2)` ``). Column names which are Python keywords (like "list", "for", "import", etc) cannot be used.

For example, if one of your columns is called `a a` and you want to sum it with `b`, your query should be `` `a a` + b ``.

La librería Pandas: DataFrames

Group By



pandas.DataFrame.groupby

```
DataFrame.groupby(by=None, axis=0, level=None, as_index=True, sort=True, group_keys=True,
squeeze=NoDefault.no_default, observed=False, dropna=True)                            [source]
```

Group DataFrame using a mapper or by a Series of columns.

A groupby operation involves some combination of splitting the object, applying a function, and combining the results. This can be used to group large amounts of data and compute operations on these groups.

Veamos esto en un jupyter notebook…

Algunos cursos de data science

- IIC2613 - Inteligencia Artificial
- IIC3633 - Sistemas recomendadores
- IIC2433 - Minería de datos
- IIC3697 - Aprendizaje profundo
- IIC2026 - Visualización de información
- IIC3733 - Visión por computador
- IMT3120 - Fundamentos matemáticos para ciencia de datos