

Python - pandas

Máster en Data Science y Big Data

Miguel Ángel Corella

mcorella@geoblink.com

Octubre 2021

Contenido

1. ¿Qué es pandas?
2. Capacidades principales
3. Referencias

¿Qué es pandas?

¿Qué es pandas?

- pandas es un módulo de Python, orientado al análisis de datos.
- Inicialmente fue creado por Wes McKinney (autor del libro “Python for Data Analysis”, referencia de este curso).
- La primera versión se publicó en 2008 y la última disponible es de Octubre de 2021 y es una de las librerías con mayor evolución y seguimiento por parte de la comunidad:
 - Más de 2.400 contribuidores
 - Más de 27.900 commits.
- Con los años se ha convertido en el estándar *de facto* para el análisis de datos en Python.

Capacidades principales

Capacidades principales (I)

- Almacenamiento y procesamiento de diferentes estructuras de datos: información tabular, información matricial, series temporales...
- Facilidad para la carga de información desde diversas fuentes: ficheros de texto, bases de datos relacionales, etc.
- Operaciones de agregación, filtrado, agrupación y ordenación sobre estructuras de datos.
- Utilidades tanto para la carga, tratamiento y limpieza de datos como para el análisis estadístico, exploratorio de datos y modelado.
- Integración directa con otras librerías como numpy (sobre la que está implementada) o scikit-learn.

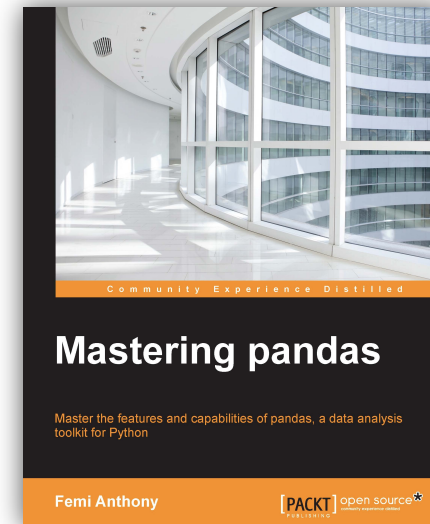
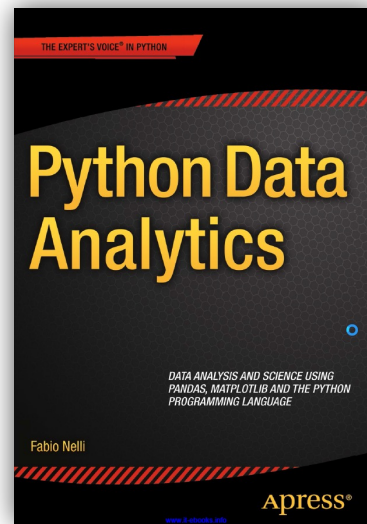
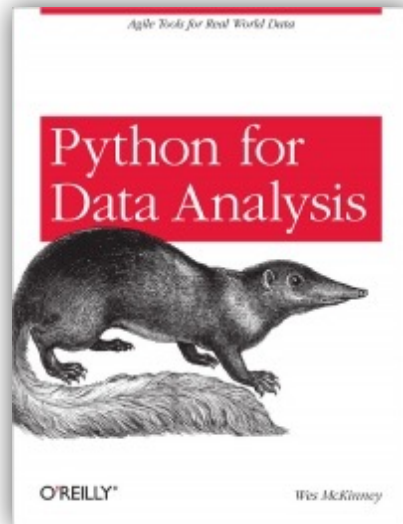
Capacidades principales (II)

- Alto rendimiento, que puede ser incluso mayor haciendo uso de Cython (que permite integrar extensiones escritas en C en programas desarrollados en Python).
- En esencia, incorpora a Python estructuras de datos y operaciones como las existentes en lenguajes de programación directamente orientados al análisis de datos como R:
 - Estructuras de datos: vectores, `data.frame`, `data.table`...
 - Operaciones: familia `apply`, agregación y agrupación, filtrado...

Referencias

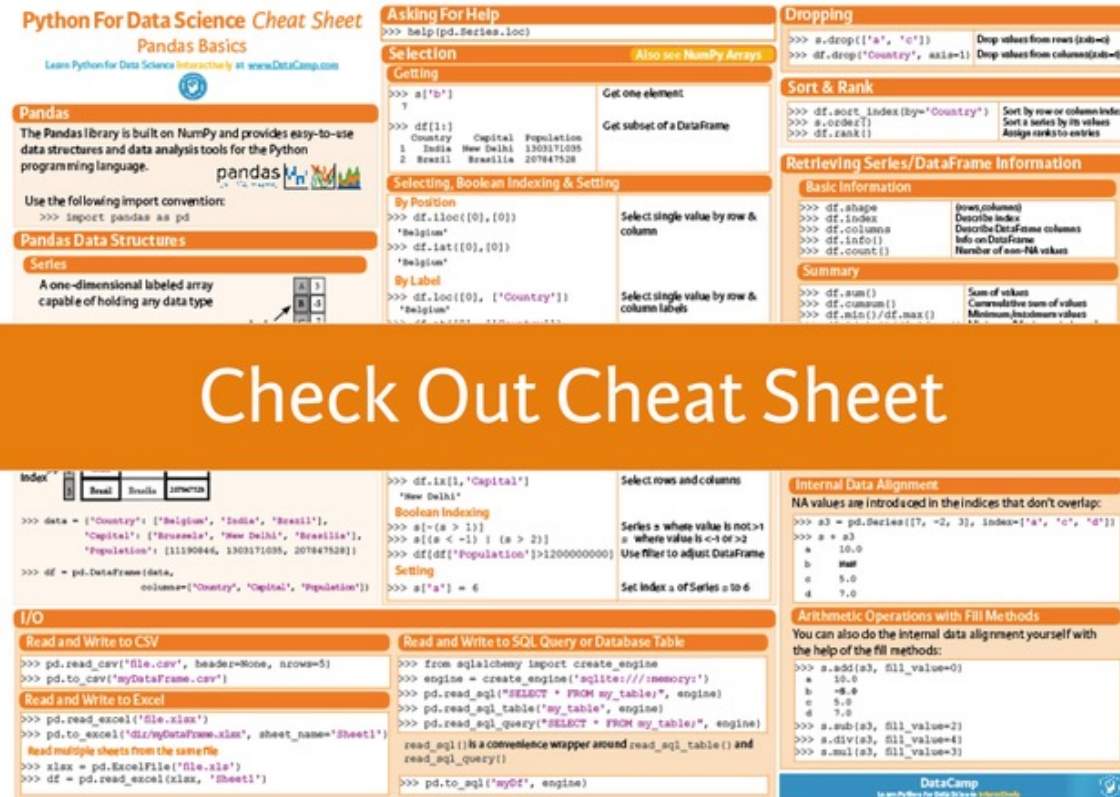
Referencias

- Página oficial de pandas:
 - <https://pandas.pydata.org/>
- Algunos libros:



Referencias

... o mucho más sencillo



The image displays a comprehensive cheat sheet for Pandas, titled "Python For Data Science Cheat Sheet: Pandas Basics". It is organized into several sections, each with code snippets and brief explanations:

- Pandas Basics:** Explains that Pandas is built on NumPy and provides easy-to-use data structures and analysis tools for Python. It includes the import convention: `>>> import pandas as pd`.
- Pandas Data Structures:** Defines a **Series** as a one-dimensional labeled array capable of holding any data type.
- Getting:** Shows how to get an element (`a["b"]`) or a subset of a DataFrame (`df[1:]`).
- Selection:** Includes **Boolean Indexing** (e.g., `df[df["Population"] > 1000000000]`) and **Setting** (e.g., `a["a"] = 6`).
- Asking For Help:** Shows `help(pd.Series.loc)`.
- Dropping:** Demonstrates dropping rows (`df.drop(["a", "c"], axis=1)`) and columns (`df.drop("Country", axis=1)`).
- Sort & Rank:** Shows sorting by row or column index (`df.sort_index(by="Country")`) and assigning ranks (`df.rank()`).
- Retrieving Series/DataFrame Information:** Includes **Basic Information** (e.g., `df.shape`, `df.index`, `df.columns`, `df.info()`, `df.count()`) and **Summary** (e.g., `df.sum()`, `df.cumsum()`, `df.min()/df.max()`).
- I/O:** Contains sections for **Read and Write to CSV**, **Read and Write to Excel**, and **Read and Write to SQL Query or Database Table** using SQLAlchemy.
- Internal Data Alignment:** Explains that NA values are introduced in indices that don't overlap and shows how to handle them with `fill_value` in arithmetic operations.

A large orange banner in the center reads "Check Out Cheat Sheet". The bottom right corner features the DataCamp logo and the text "Learn Python For Data Science Interactively at www.DataCamp.com".

<http://datacamp-community-prod.s3.amazonaws.com/dbed353d-2757-4617-8206-8767ab379ab3>



Afi Escuela

© 2021 Afi Escuela. Todos los derechos reservados.