



Pandas

Minicurso de introdução ao Python e seu ecossistema científico

João Pedro dos S. Rocha

UESC-DCET

O que levar para casa?

- O Pandas é construído por cima do numpy e estende suas capacidades facilitando a manipulação de dados heterogêneos e empacotando diversas funcionalidades.
- As funcionalidades do Pandas giram em torno do objeto DataFrame que é composto de 2 componentes. Isso nos permite acessar os dados de várias maneiras.
- O Pandas foi criado para trabalhar bem com outras ferramentas como o SQL, por isso ele funciona bem com o modelo Tidy Data.

O que é o Pandas?

- O pandas é a biblioteca do Python que trás funcionalidades que facilitam significativamente o trabalho com dados tabulares.
- Ela facilita a manipulação de dados heterogêneos, e o uso de etiquetas.
- Nos dá acesso a uma grande gamma de ferramentas de ferramentas que facilita todo o processo de análise de dados
- Carrega a herança do SQL e outras ferramentas estatísticas (como o R) foi modelado usando o Tidy Data

Por que usar o Pandas?

- Simples para manipular dados heterogêneos
- Facilidade para lidar com dados ausentes
- Fácil de usar os groupby e outras operações (estilo SQL)
- Alinhamento automático de dados
- Excelente suporte para séries temporais

O modelo do Pandas

The diagram illustrates the structure of a Pandas DataFrame. It features a table with 10 columns and 7 rows. Annotations include: 'Column names' pointing to the header row; 'Columns axis=1' pointing to the column headers; 'Index label' pointing to the index column; 'Index axis=0' pointing to the index values; 'Missing value' pointing to a 'NaN' cell in the 'Number' column; and 'Data' pointing to the body of the table.

Column names									
	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0.0	PG	25.0	6-2	180.0	Texas	7730337.0
1	John Holland	Boston Celtics	30.0	SG	27.0	6-5	205.0	Boston Uniersity	NaN
2	Jonas Jerebko	Boston Celtics	8.0	PF	29.0	6-10	231.0	NaN	5000000.0
3	Jordan Mickey	Boston Celtics	NaN	PF	21.0	6-8	235.0	LSU	1170960.0
4	Terry Rozier	Boston Celtics	12.0	PG	22.0	6-2	190.0	Louisville	1824360.0
5	Jared Sullinger	Boston Celtics	7.0	C	NaN	6-9	260.0	Ohio State	2569260.0
6	Evan Turner	Boston Celtics	11.0	SG	27.0	6-7	220.0	Ohio State	3425510.0

fonte: [geeksforgeeks.org](https://www.geeksforgeeks.org/)

Tidy Data

- Muito importante !!!
- Wikham 2014

country	year	cases	population
Afghanistan	1999	18215	19987071
Afghanistan	2000	18666	20095360
Brazil	1999	31737	172006362
Brazil	2000	80488	174004898
China	1999	210258	1272015272
China	2000	213666	1280008583

variables

country	year	cases	population
Afghanistan	1999	18215	19987071
Afghanistan	2000	18666	20095360
Brazil	1999	31737	172006362
Brazil	2000	80488	174004898
China	1999	210258	1272015272
China	2000	213666	1280008583

observations

country	year	cases	population
Afghanistan	1999	18215	19987071
Afghanistan	2000	18666	20095360
Brazil	1999	31737	172006362
Brazil	2000	80488	174004898
China	1999	210258	1272015272
China	2000	213666	1280008583

values

fonte: garrettgman.github.io

Recapitulando ...

- O Pandas é construído por cima do numpy e estende suas capacidades facilitando a manipulação de dados heterogêneos e empacotando diversas funcionalidades.
- As funcionalidades do Pandas giram em torno do objeto DataFrame que é composto de 2 componentes. Isso nos permite acessar os dados de várias maneiras.
- O Pandas foi criado para trabalhar bem com outras ferramentas como o SQL, por isso ele funciona bem com o modelo Tidy Data.