

# Pandas

---

Advanced Institute for Artificial Intelligence

<https://advancedinstitute.ai>

- ☐ Pandas
- ☐ Estruturas
- ☐ Operações

- Biblioteca Python de código aberto para análise de dados
- Pandas é composto por dois tipos de estruturas de dados:
  - Series
  - DataFrame

- ☐ Uma série é um objeto unidimensional semelhante a uma matriz, lista ou coluna em uma tabela.
- ☐ Cada item possui um índice, por padrão o índice é de 0 a N
- ☐ Alternativamente, você pode especificar um índice para a série.

- ☐ O construtor Series também pode converter um dicionário, usando as chaves do dicionário como índice.
- ☐ Você pode usar o índice para selecionar itens específicos da série
- ☐ É possível utilizar uma condição, que ao ser aplicada ao índice, retorna os elementos que atendem a condição
- ☐ Operações matemáticas podem ser feitas usando escalares e funções.

## DataFrame

- Estrutura de dados tabular composta de linhas e colunas
  - planilha, tabela de banco de dados, etc
  - Um Dataframe pode ser também um grupo de objetos series

## Montando um Dataframe

- ❑ Para criar um DataFrame a partir de estruturas de dados Python comuns, podemos passar um dicionário de listas para o construtor DataFrame.
- ❑ Parâmetro column permite indicar a ordem das colunas

## Montando um Dataframe

- ❑ Para criar um DataFrame a partir de estruturas de dados Python comuns, podemos passar um dicionário de listas para o construtor DataFrame.
- ❑ Parâmetro column permite indicar a ordem das colunas
- ❑ É possível criar um Dataframe a partir de um CSV
  - `fromcsv = pd.read_csv('teste.csv')`
- ❑ É possível criar um Dataframe a partir de uma URL
  - `from_url = pd.read_table(url)`



- Info() : Retorna informações diversas
  - número de linhas, colunas, quantidade de espaço ocupado e tipo de dado
- Describe() : Retorna informações estatísticas a respeito dos dados de cada coluna

- Definindo uma coluna do dataframe como índice
  - `users.set_index('user_id', inplace = True)`
- `inplace=true` indica que a própria referência deve ser alterada, senão um novo dataframe é retornado

## Seleção de linhas e colunas

☐ A seleção de uma coluna retorna um objeto series

- `display(from_url[['rank', 'price']].head())`
- `print(from_url[['rank', 'price']].head())`

☐ Selecionando linhas

- `display(users[users.age > 25].head(3))`

Seleção de linhas e colunas loc, iloc e ix:

- ❑ `.loc []` funciona nos rótulos do seu índice. Isso significa que, se você fornecer o `loc [2]`, procurará os valores do seu DataFrame com um índice rotulado 2.
- ❑ `.iloc []` trabalha nas posições em seu índice. Isso significa que, se você fornecer `iloc [2]`, procurará os valores do seu DataFrame que estão no índice '2'.
- ❑ `.ix []` é um caso mais complexo: quando o índice é baseado em números inteiros, você passa um rótulo para `.ix []`.

## Exemplos

```
df = pd.DataFrame(data=np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9]]), index= [2
```

```
# Pass '2' to 'loc'  
print(df.loc[2])
```

```
# Pass '2' to 'iloc'  
print(df.iloc[2])
```

```
# Pass '2' to 'ix'  
print(df.ix[2])
```

## Adicionando colunas

- basta utilizar um nome de coluna inexistente e passar um objeto series como parâmetro