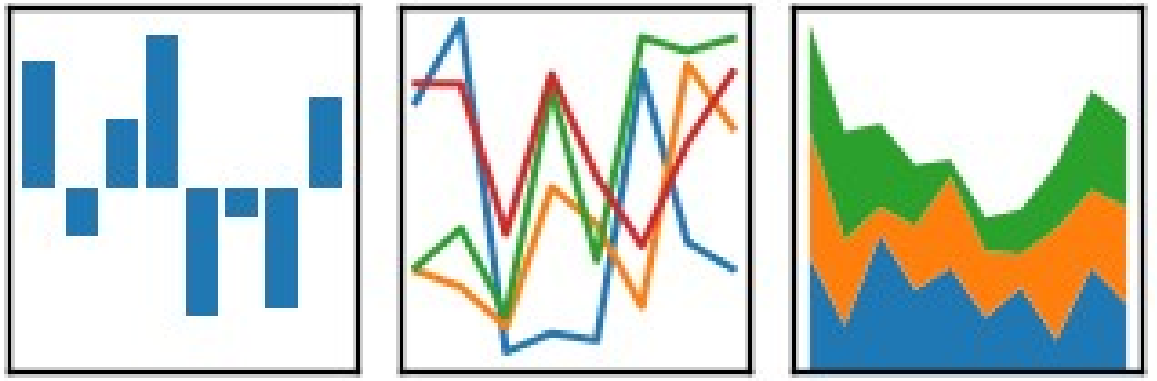
Pandas - Introdução

Luiz Celso Gomes-Jr

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



- Used for structured data operations and manipulations. It is extensively used for data munging and preparation.
- Have been instrumental in boosting Python’s usage in data scientist community.

DataFrame

- DataFrames são dados estruturados em formato de planilha.

```
import pandas as pd

df = pd.DataFrame({'nome': ['Celso', 'Josie', 'Luiz', 'Helen'],
                   'idade': ['23', '32', '11', '18'],
                   'genero': ['M', 'F', 'M', 'F'],
                  }).set_index('nome')

df
```

nome	idade	genero
Celso	23	M
Josie	32	F
Luiz	11	M
Helen	18	F

Lendo dados de um arquivo

- É muito fácil importar dados em CSV usando o Pandas

```
# lê o arquivo CSV
df = pd.read_csv('../data/aluguel.csv')

# mostra o conteúdo do DataFrame
df.head()
```

	codigo	endereco	quartos	suite	area	vaga	aluguel	condominio	data
0	34	Rua Desembargador Westphalen	2	0	90	0	900	371	11/10/17
1	167	Rua Jose Loureiro	2	0	64	0	650	428	15/07/17
2	6784	Rua Jose Loureiro	2	0	81	0	1100	400	23/08/17
3	82	Rua Lourenço Pinto	2	0	50	0	1350	300	19/09/17
4	2970	Rua Lourenço Pinto	2	0	63	0	1300	300	05/08/17

Lendo dados de um arquivo

- É também fácil importar dados em **Excel** usando o Pandas

```
df = pd.read_excel(open("../data/UC_modelos.xlsx","rb"),
                   sheet_name = "exemplos")
df.head(2)
```

	Unidade de Conservação	Unnamed: 1	Ano de fundação/ inauguração	Área (Há)	Regional	Legislação de criação	Visitação pública com autorização prévia	Visit pú perm
0	Parque Nacional do Itatiaia	NaN	1937.0	30000	MG, RJ	NaN	NÃO	
1	Parque Nacional das Sempre Vivas	NaN	2001.0	124156	MG	NaN	NÃO	

Selecionando colunas do DataFrame

É possível criar novos DataFrames com subconjuntos de colunas do DataFrame original. Abaixo um DataFrame chamado df_small é criado com o conteúdo das colunas aluguel e condominio.

```
df.head(3)
```

	codigo	endereco	quartos	suite	area	vaga	aluguel	condominio	data
0	34	Rua Desembargador Westphalen	2	0	90	0	900	371	11/10/17
1	167	Rua Jose Loureiro	2	0	64	0	650	428	15/07/17
2	6784	Rua Jose Loureiro	2	0	81	0	1100	400	23/08/17

```
df_small = df[['aluguel', 'condominio']]

df_small.head(2)
```

	aluguel	condominio
0	900	371
1	650	428

Selecionando linhas do DataFrame

Também é possível selecionar subconjuntos das linhas do DataFrame. Abaixo selecionamos apenas as linhas representando apartamentos de área maior que $50m^2$.

```
df[df.area > 50]
```

	codigo	endereco	quartos	suite	area	vaga	aluguel	condominio	data
0	34	Rua Desembargador Westphalen	2	0	90	0	900	371	11/10/17
1	167	Rua Jose Loureiro	2	0	64	0	650	428	15/07/17
2	6784	Rua Jose Loureiro	2	0	81	0	1100	400	23/08/17
4	2970	Rua Lourenço Pinto	2	0	63	0	1300	300	05/08/17
5	34197	Alameda Doutor Muricy	2	0	80	1	900	410	23/10/17
6	5072	Alameda Doutor Muricy	2	0	84	0	1100	382	02/09/17

Novas colunas

É possível criar novas colunas a partir de outras colunas do DataFrame. Abaixo criamos uma coluna chamada total, contendo o valor do aluguel somado ao condomínio.

```
df['total'] = df['aluguel'] + df['condominio']

df
```

	codigo	endereco	quartos	suite	area	vaga	aluguel	condominio	data	total
0	34	Rua Desembargador Westphalen	2	0	90	0	900	371	11/10/17	1271
1	167	Rua Jose Loureiro	2	0	64	0	650	428	15/07/17	1078
2	6784	Rua Jose Loureiro	2	0	81	0	1100	400	23/08/17	1500
3	82	Rua Lourenço Pinto	2	0	50	0	1350	300	19/09/17	1650
4	2970	Rua Lourenço Pinto	2	0	63	0	1300	300	05/08/17	1600
5	34197	Alameda Doutor Muricy	2	0	80	1	900	410	23/10/17	1310
6	5072	Alameda Doutor Muricy	2	0	84	0	1100	382	02/09/17	1482

Exercícios!

- Revise o conteúdo e faça os exercícios do notebook 03a-Pandas_Introdução.ipynb

Índices

- Usados para acessar valores específicos dentro de um DataFrame.
- Por exemplo, usando índices podemos selecionar valores de determinadas linhas ou colunas de um DataFrame.

Índices – Definindo o índice de uma DF

```
df.head(3)
```

	codigo	endereco	quartos	suite	area	vaga	aluguel	condominio	data
0	34	Rua Desembargador Westphalen	2	0	90	0	900	371	11/10/17
1	167	Rua Jose Loureiro	2	0	64	0	650	428	15/07/17
2	6784	Rua Jose Loureiro	2	0	81	0	1100	400	23/08/17

```
df = df.set_index('codigo')

df.head(3)
```

	endereco	quartos	suite	area	vaga	aluguel	condominio	data
codigo								
34	Rua Desembargador Westphalen	2	0	90	0	900	371	11/10/17
167	Rua Jose Loureiro	2	0	64	0	650	428	15/07/17
6784	Rua Jose Loureiro	2	0	81	0	1100	400	23/08/17

Índices – Ordenação

```
df = df.sort_index()

df
```

codigo	endereco	quartos	suite	area	vaga	aluguel	condominio	data
24	Rua Desembargador Westphalen	1	0	60	1	800	120	30/09/17
34	Rua Desembargador Westphalen	2	0	90	0	900	371	11/10/17
74	Avenida Visconde de Guarapuava	2	1	132	1	1800	520	12/10/17
80	Rua Desembargador Westphalen	1	0	80	1	900	350	12/08/17
82	Rua Lourenço Pinto	2	0	50	0	1350	300	19/09/17
167	Rua Jose Loureiro	2	0	64	0	650	428	15/07/17
469	Rua Desembargador Westphalen	1	0	30	0	550	210	03/07/17
568	Rua Alferes Poli	1	0	43	0	600	330	12/08/17
2381	Rua Rockefeller	2	0	54	0	900	240	19/09/17

Índices – Tipo e conteúdo

```
df
```

codigo	endereco	quartos	suite	area	vaga	aluguel	condominio	data
24	Rua Desembargador Westphalen	1	0	60	1	800	120	30/09/17
34	Rua Desembargador Westphalen	2	0	90	0	900	371	11/10/17
74	Avenida Visconde de Guarapuava	2	1	132	1	1800	520	12/10/17
80	Rua Desembargador Westphalen	1	0	80	1	900	350	12/08/17
82	Rua Lourenço Pinto	2	0	50	0	1350	300	19/09/17

```
indice = df.index

print(type(indice))
indice
```

```
<class 'pandas.core.indexes.numeric.Int64Index'>

Int64Index([24, 34, 74, 80, 82], dtype='int64', name='codigo')
```

Índices – Colunas

Colunas também são índices! Veja abaixo como obter a lista de colunas e preste atenção no tipo (classe).

```
colunas = df.columns

print(type(colunas))
colunas
```

```
<class 'pandas.core.indexes.base.Index'>

Index(['endereco', 'quartos', 'suite', 'area', 'vaga', 'aluguel',
'condominio',
       'data'],
      dtype='object')
```

Índices – Referenciando valores

Como tanto as linhas quanto as colunas são índices, podemos usar os dois para acessar valores no DataFrame. Usando o método **loc** podemos especificar valores de índice para obter dados do DataFrame. Abaixo obtermos o valor da linha de índice 82 e coluna area.

```
df
```

codigo	endereco	quartos	suite	area	vaga	aluguel	condominio	data
24	Rua Desembargador Westphalen	1	0	60	1	800	120	30/09/17
34	Rua Desembargador Westphalen	2	0	90	0	900	371	11/10/17
74	Avenida Visconde de Guarapuava	2	1	132	1	1800	520	12/10/17
80	Rua Desembargador Westphalen	1	0	80	1	900	350	12/08/17
82	Rua Lourenço Pinto	2	0	50	0	1350	300	19/09/17

```
df.loc[82,'area']
```

```
50
```

Índices – Faixas de valores

Podemos também especificar faixas de valores usando o operador '**:**'

codigo	endereco	quartos	suite	area	vaga	aluguel	condominio	data
24	Rua Desembargador Westphalen	1	0	60	1	800	120	30/09/17
34	Rua Desembargador Westphalen	2	0	90	0	900	371	11/10/17
74	Avenida Visconde de Guarapuava	2	1	132	1	1800	520	12/10/17
80	Rua Desembargador Westphalen	1	0	80	1	900	350	12/08/17
82	Rua Lourenço Pinto	2	0	50	0	1350	300	19/09/17

```
df.loc[50:80,'area':'aluguel']
```

codigo	area	vaga	aluguel
74	132	1	1800
80	80	1	900

Índices – Seleção posicional

Para desconsiderar os índices e obter valores pelas suas posições no DataFrame, use o método **iloc**.

codigo	endereco	quartos	suite	area	vaga	aluguel	condominio	data
24	Rua Desembargador Westphalen	1	0	60	1	800	120	30/09/17
34	Rua Desembargador Westphalen	2	0	90	0	900	371	11/10/17
74	Avenida Visconde de Guarapuava	2	1	132	1	1800	520	12/10/17
80	Rua Desembargador Westphalen	1	0	80	1	900	350	12/08/17
82	Rua Lourenço Pinto	2	0	50	0	1350	300	19/09/17

```
df.iloc[3,3]
```

```
80
```

Índices – Seleção lógica

Outra forma de selecionar valores é especificar as posições de interesse usando um vetor contendo True para as posições desejadas e False para as posições indesejadas.

codigo	endereco	quartos	suite	area	vaga	aluguel	condominio	data
24	Rua Desembargador Westphalen	1	0	60	1	800	120	30/09/17
34	Rua Desembargador Westphalen	2	0	90	0	900	371	11/10/17
74	Avenida Visconde de Guarapuava	2	1	132	1	1800	520	12/10/17
80	Rua Desembargador Westphalen	1	0	80	1	900	350	12/08/17
82	Rua Lourenço Pinto	2	0	50	0	1350	300	19/09/17

```
selecionar = [True, False, True, False, False]
df[selecionar]
```

codigo	endereco	quartos	suite	area	vaga	aluguel	condominio	data
24	Rua Desembargador Westphalen	1	0	60	1	800	120	30/09/17
74	Avenida Visconde de Guarapuava	2	1	132	1	1800	520	12/10/17

Índices – Seleção lógica

Podemos gerar vetores lógicos a partir de comparações matemáticas, como > (maior), >= (maior ou igual), == (igual), != (diferente)

codigo	endereco	quartos	suite	area	vaga	aluguel	condominio	data
24	Rua Desembargador Westphalen	1	0	60	1	800	120	30/09/17
34	Rua Desembargador Westphalen	2	0	90	0	900	371	11/10/17
74	Avenida Visconde de Guarapuava	2	1	132	1	1800	520	12/10/17
80	Rua Desembargador Westphalen	1	0	80	1	900	350	12/08/17
82	Rua Lourenço Pinto	2	0	50	0	1350	300	19/09/17

```
selecionar = df['aluguel'] <= 900
selecionar
```

```
codigo
24     True
34     True
74    False
80     True
82    False
Name: aluguel, dtype: bool
```

Índices – Seleção lógica

codigo	endereco	quartos	suite	area	vaga	aluguel	condominio	data
24	Rua Desembargador Westphalen	1	0	60	1	800	120	30/09/17
34	Rua Desembargador Westphalen	2	0	90	0	900	371	11/10/17
74	Avenida Visconde de Guarapuava	2	1	132	1	1800	520	12/10/17
80	Rua Desembargador Westphalen	1	0	80	1	900	350	12/08/17
82	Rua Lourenço Pinto	2	0	50	0	1350	300	19/09/17

```
selecionar = df['aluguel'] != 900
df[selecionar]
```

codigo	endereco	quartos	suite	area	vaga	aluguel	condominio	data
24	Rua Desembargador Westphalen	1	0	60	1	800	120	30/09/17
74	Avenida Visconde de Guarapuava	2	1	132	1	1800	520	12/10/17
82	Rua Lourenço Pinto	2	0	50	0	1350	300	19/09/17

Índices – Seleção lógica

```
selecionar = df['aluguel'] <= 900
df[selecionar]
```

codigo	endereco	quartos	suite	area	vaga	aluguel	condominio	data
24	Rua Desembargador Westphalen	1	0	60	1	800	120	30/09/17
34	Rua Desembargador Westphalen	2	0	90	0	900	371	11/10/17
80	Rua Desembargador Westphalen	1	0	80	1	900	350	12/08/17

```
df[df['aluguel'] <= 900]
```

codigo	endereco	quartos	suite	area	vaga	aluguel	condominio	data
24	Rua Desembargador Westphalen	1	0	60	1	800	120	30/09/17
34	Rua Desembargador Westphalen	2	0	90	0	900	371	11/10/17
80	Rua Desembargador Westphalen	1	0	80	1	900	350	12/08/17

Índices – Reiniciar

Para reiniciar o índice, uso o comando **reset_index**. Este comando copia o índice anterior para uma coluna e cria um novo índice sequencial.

```
df = df.reset_index()
df
```

	codigo	endereco	quartos	suite	area	vaga	aluguel	condominio	data
0	24	Rua Desembargador Westphalen	1	0	60	1	800	120	30/09/17
1	34	Rua Desembargador Westphalen	2	0	90	0	900	371	11/10/17
2	74	Avenida Visconde de Guarapuava	2	1	132	1	1800	520	12/10/17
3	80	Rua Desembargador Westphalen	1	0	80	1	900	350	12/08/17
4	82	Rua Lourenço Pinto	2	0	50	0	1350	300	19/09/17

Exercícios!

- Revise o conteúdo e faça os exercícios do notebook (menos a parte de MultiIndex): 03b-Pandas_Índices e Seleção de Valores.ipynb

Índices hierárquicos (MultiIndex)

- Índices de múltiplos níveis são chamados MultiIndex.
- São gerados como resultado de operações de processamento de dados como groupby.

MultiIndex – Definição

```
df_multi_index = df.set_index(['quartos', 'vaga'])
df_multi_index
```

quartos	vaga	codigo	endereco	suite	area	aluguel	condominio	data
1	1	24	Rua Desembargador Westphalen	0	60	800	120	30/09/17
2	0	34	Rua Desembargador Westphalen	0	90	900	371	11/10/17
	1	74	Avenida Visconde de Guarapuava	1	132	1800	520	12/10/17
1	1	80	Rua Desembargador Westphalen	0	80	900	350	12/08/17
2	0	82	Rua Lourenço Pinto	0	50	1350	300	19/09/17

MultiIndex – Ordenação

```
df_multi_index.sort_index()
```

		codigo	endereco	suite	area	aluguel	condominio	data
quartos	vaga							
1	1	24	Rua Desembargador Westphalen	0	60	800	120	30/09/17
	1	80	Rua Desembargador Westphalen	0	80	900	350	12/08/17
2	0	34	Rua Desembargador Westphalen	0	90	900	371	11/10/17
	0	82	Rua Lourenço Pinto	0	50	1350	300	19/09/17
	1	74	Avenida Visconde de Guarapuava	1	132	1800	520	12/10/17

MultiIndex – Referenciando valores

O método **loc**, para seleção de dados, também pode ser usado em um índice multinível. Os valores de cada nível devem ser especificados numa tupla (entre parêntesis).

quartos	vaga	codigo	endereco	suite	area	aluguel	condominio	data
1	1	24	Rua Desembargador Westphalen	0	60	800	120	30/09/17
	1	80	Rua Desembargador Westphalen	0	80	900	350	12/08/17
2	0	34	Rua Desembargador Westphalen	0	90	900	371	11/10/17
	0	82	Rua Lourenço Pinto	0	50	1350	300	19/09/17
	1	74	Avenida Visconde de Guarapuava	1	132	1800	520	12/10/17

```
df_multi_index.loc[(2,1), 'aluguel']
```

```
quartos  vaga
2        1       1800
Name: aluguel, dtype: int64
```