

Pandas

primeiros passos de um cientista de dados

Storytelling - Imobiliária

Aluga Rápido

Atores

- Estagiário de Ciência de Dados (você)
- Engenheiro de IA
- Desenvolvedor Django/Flask

Introdução à Biblioteca Pandas

O que é e por que usar Pandas?

- Você acabou de ser contratado como estagiário na imobiliária mais badalada da cidade. O chefe te entrega uma planilha Excel com milhares de linhas de clientes e propriedades. É para surtar? Claro que não! Aqui entra o Pandas, uma biblioteca Python de código aberto que transforma o caos em uma sinfonia de dados!

DataFrames - Seu Novo Melhor Amigo

- Transformando Dados em Informação
- Com Pandas, você pode importar dados de várias fontes para um DataFrame, que é basicamente uma super planilha do Excel no mundo Python. Imagine filtrar, ordenar e transformar todos os dados da imobiliária com apenas algumas linhas de código. Vai ser moleza encontrar o cliente perfeito para aquele imóvel de luxo!

O Poder das Operações

- Filtrar, Ordenar e Agregar
- Pandas te permite realizar operações incríveis como filtrar os dados para encontrar os clientes VIP, ordenar as propriedades pelo preço e agregar informações para descobrir os bairros mais procurados. É como ter uma lupa mágica para ver tudo de forma clara e organizada.

Aplicações do Pandas

- De Ciência de Dados à Análise de Negócios
- Pandas não é só para cientistas de dados em laboratórios secretos. Na imobiliária, ele ajuda em tudo: da análise de mercado para identificar tendências de preços à segmentação de clientes para campanhas de marketing. É a ferramenta que todo corretor esperto gostaria de ter

Resumo

- Pandas é uma ferramenta poderosa que transforma como você trabalha com dados. Seja para criar relatórios detalhados ou visualizar informações complexas de maneira simples, é a melhor amiga dos analistas, cientistas e engenheiros de dados. Imagine se o Excel fosse um super-herói. É isso.

Agenda (*competências*)

- Explorar as características de uma base
- Realizar análises exploratórias com diferentes métodos
- Lidar com valores nulos
- Remover registros inconsistentes
- Aplicar filtros
- Criar colunas
- Entre outras

Pré-requisitos

- Fundamentos de Python
- Fundamentos de Data Science
- Fundamentos de IA
- Fundamentos de Versionamento

Profissionais

- Cientista de Dados (você)
- Engenheiro de Machine Learning
- Desenvolvedor Full Stack Django/Flask

O problema

- Atualmente a imobiliária Aluga Rápido, necessita de um cientista de dados para dar apoio a duas equipes “IA” e “Dev”.
- Você foi contratado e disponibilizaram:
 - Base de dados de imóveis (preços dos aluguéis)
 - Por ser Home Office de Brasília para uma empresa do Rio de Janeiro:
 - Utilizam o Trello com as demandas (tarefas de trabalho)

Demandas

- Informações do Projeto
 - Base de Dados
- A fazer
 - Importar e conhecer a base de dados
 - **Time de ML (roxa)**
 - Análise exploratória dos dados
 - Tratar valores nulos
 - Remover registros inconsistentes
 - Aplicar filtros
 - **Time do Dev (verde)**
 - Criação de colunas numéricas
 - Criação de colunas categóricas
- Em Andamento
- Concluído

Trello

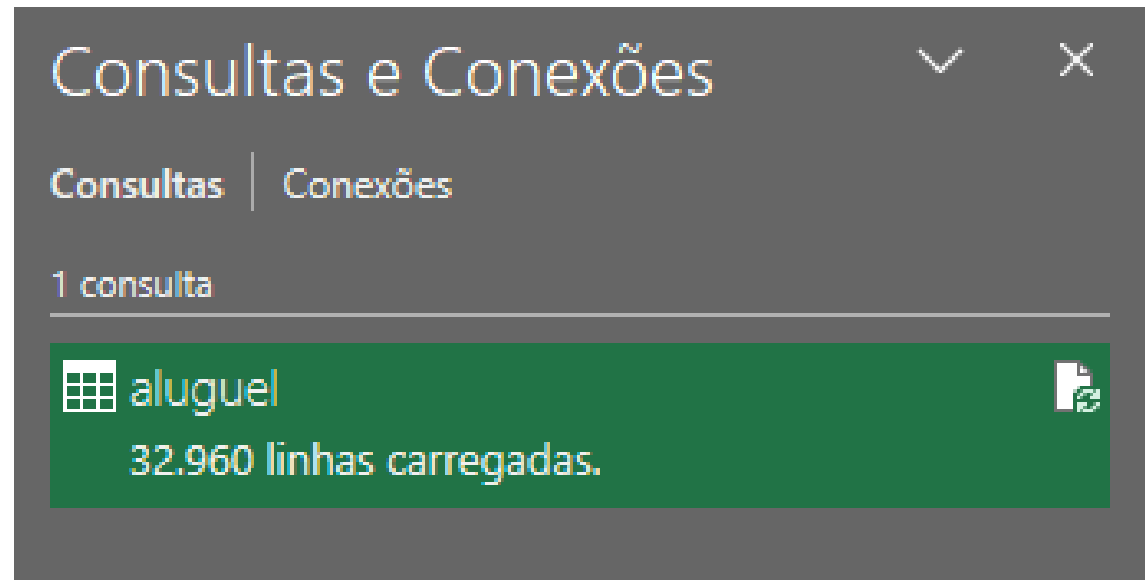
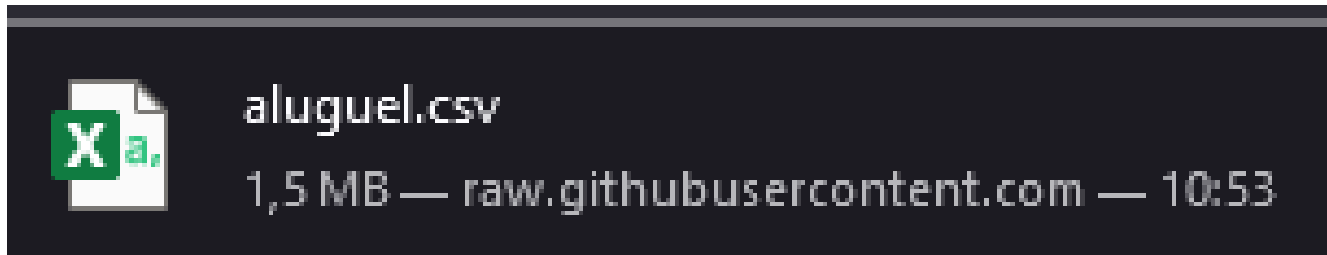
- Nós fomos contratados como cientistas de dados de uma empresa imobiliária. Nossa principal função é dar suporte as demandas do time de Machine Learning e do time de Desenvolvimento dessa empresa.
- Para atendermos essas demandas, foi disponibilizado um board no Trello, com as etapas e tarefas do projeto que devemos realizar. Também foi disponibilizada uma base de dados, que utilizaremos para desenvolver o projeto.

Base de Dados

- A base de dados que vamos utilizar para desenvolver o projeto é uma base com dados de diferentes tipos de imóveis do Rio de Janeiro, como apartamento, casas, comércios, dentre outros.
- Nessa base, nós vamos encontrar os valores dos aluguéis de cada imóvel, condomínio, IPTU e também suas características, como: quantidade de quartos, suítes, vagas de garagem, etc.

Base de Dados

- Vamos abrir o nossa base de dados no Excel



	A	B	C	D	E	F	G	H	I
1	Tipo	Bairro	Quartos	Vagas	Suites	Area	Valor	Condominio	IPTU
2	Quitinete	Copacabana	1	0	0	40	1700	500	60
3	Casa	Jardim Botânico	2	0	1	100	7000		
4	Conjunto Comercial/Sala	Barra da Tijuca	0	4	0	150	5200	4020	1111
5	Apartamento	Centro	1	0	0	15	800	390	20
6	Apartamento	Higienópolis	1	0	0	48	800	230	
7	Apartamento	Vista Alegre	3	1	0	70	1200		
8	Apartamento	Cachambi	2	0	0	50	1300	301	17
9	Casa de Condomínio	Barra da Tijuca	5	4	5	750	22000		
10	Casa de Condomínio	Ramos	2	2	0	65	1000		
11	Conjunto Comercial/Sala	Centro	0	3	0	695	35000	19193	3030
12	Apartamento	Centro	1	0	0	36	1200		
13	Apartamento	Grajaú	2	1	0	70	1500	642	74
14	Apartamento	Linha de Macaé	2	1	1	60	1500	455	14

> Fundamentos Pandas



Quadro ▾



Filtros

RP

Compartilhar



Informações do Projeto

O Problema



O DataSet



1

+ Adicionar um cartão



A Fazer

Importar e conhecer a base de dados

Análise exploratório dos dados

Tratar valores nulos

Remover registros inconsistentes

Aplicar filtros

+ Adicionar um cartão



Desenvolvimento

não há

+ Adicionar um cartão



Concluído

não há

+ Adicionar um cartão



+ Adicionar outra lista

INOVA TECH

Execução

SENAI
PELO FUTURO DO TRABALHO

Realização

efapdf
Fundação de Amparo à
Pesquisa do Estado de PernambucoSecretaria de
Ciência, Tecnologia
e Inovação

Ambiente do Curso

- Nosso curso será desenvolvido em Notebooks Jupyter
- Você deve instalar o Visual Studio Code
- Deve instalar Jupyter Notebook no VS Code
- Os códigos diariamente devem ser enviados para o github do aluno disponibilizado para o professor diariamente.
- No README.md deve ter um diário de bordo do notebook
 - O que faz o notebook
 - O que mais aprendeu na aula?
 - O que não entendeu e precisa de revisão?
 - Pontos de melhoria para a aula?

Notebook Inicial

- Entre no github do professor e baixar o notebook inicial.
- Branch : Master
 - <https://github.com/romulosilvestre/semanadatascience>

Hora um

Importar o Pandas: O módulo pandas é uma biblioteca de software escrita para a linguagem de programação Python para manipulação e análise de dados. Ela oferece estruturas de dados e operações para manipular tabelas numéricas e séries temporais.

Atribuir um alias (pd): Ao importar a biblioteca pandas, atribuímos a ela o alias pd. Isso é uma convenção comum entre programadores de Python para tornar o código mais conciso e legível. Em vez de ter que escrever pandas cada vez que você usa uma função ou classe da biblioteca, você pode simplesmente escrever pd.

SEMANDADC

> .venv

horaum.ipynb

intro.ipynb

README.md

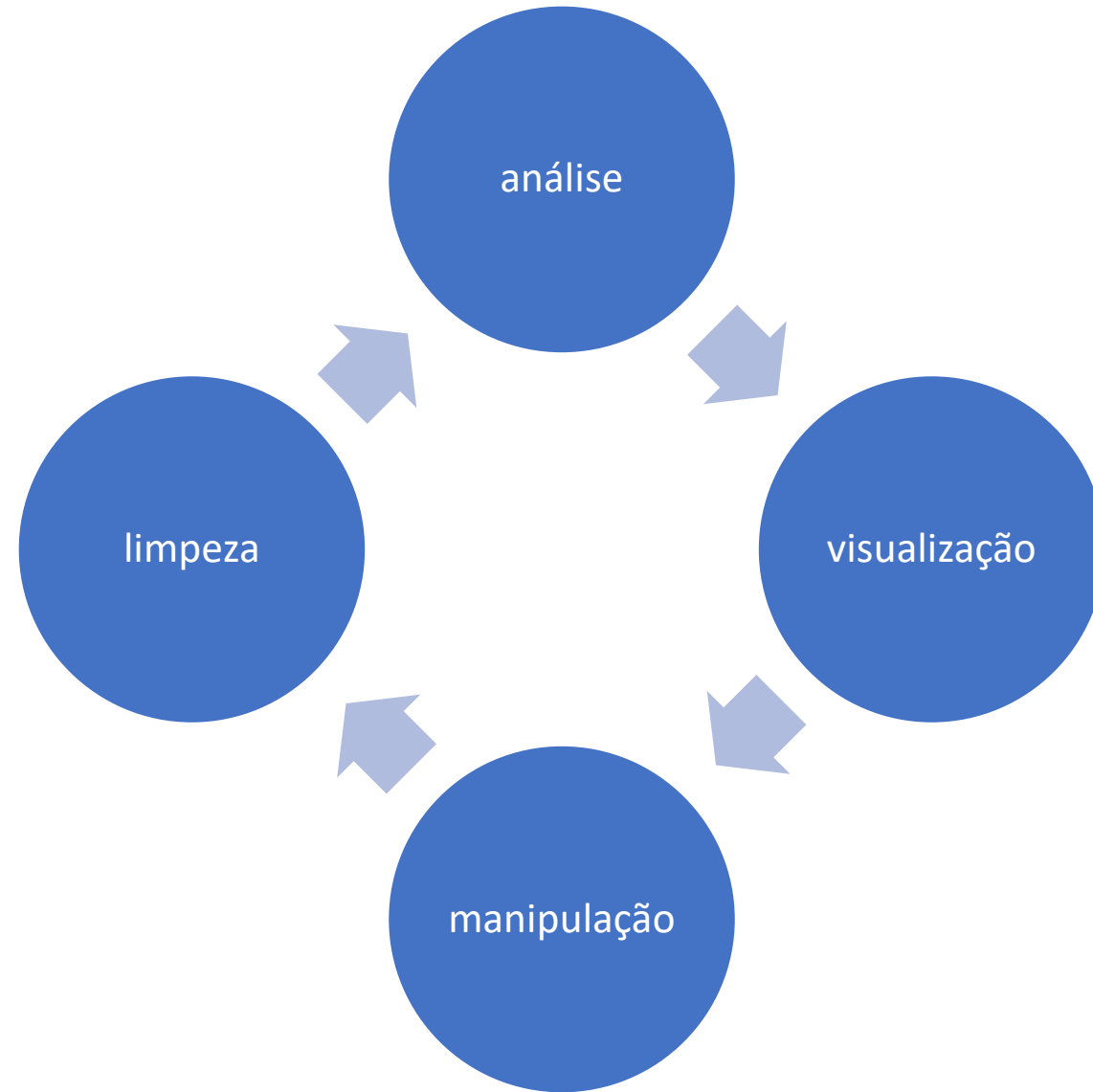
horaum.ipynb > Imobiliária Aluga Rápido

+ Code + Markdown | ▶ Run All ☰ Clear All Outputs | ☰ Outline ...

Imobiliária Aluga Rápido

Pandas e seus amigos

- Pandas – biblioteca de manipulação de dados
- NumPy – biblioteca que facilita a realização de cálculos científicos
- Scikit-Learn – biblioteca padrão para machine learn




```
(.venv) C:\Users\User\OneDrive\Área de Trabalho\semandadc>pip install pandas
```

```
Collecting pandas
```

```
  Downloading pandas-2.2.2-cp312-cp312-win_amd64.whl.metadata (19 kB)
```

```
Collecting numpy>=1.26.0 (from pandas)
```

```
  Downloading numpy-2.0.0-cp312-cp312-win_amd64.whl.metadata (60 kB)
```

```
60.9/60.9 kB 1.6 MB/s eta 0:00:00
```

```
Collecting python-dateutil>=2.8.2 (from pandas)
```

```
  Using cached python_dateutil-2.9.0.post0-py2.py3-none-any.whl.metadata (8.4 kB)
```

```
Collecting pytz>=2020.1 (from pandas)
```

```
  Using cached pytz-2024.1-py2.py3-none-any.whl.metadata (22 kB)
```

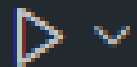
```
Collecting tzdata>=2022.7 (from pandas)
```

```
  Using cached tzdata-2024.1-py2.py3-none-any.whl.metadata (1.4 kB)
```

```
Collecting six>=1.5 (from python-dateutil>=2.8.2->pandas)
```

Imobiliária Aluga Rápido

Importando Dados



```
import pandas as pd
```

[1]



1.2s

+ Code



File

Edit

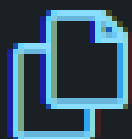
Selection

View

Go

Run

...

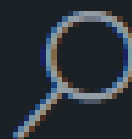


EXPLORER

...



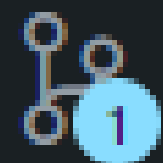
SEMANDADC



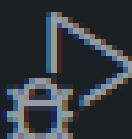
venv



dataset



1



horaum.ipynb

U

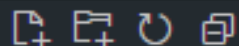


intro.ipynb



README.md

SEMANDADC



venv

Include

Lib

Scripts

share

.gitignore

pyenv.cfg

dataset

aluguel.csv

U

horaum.ipynb



U

intro.ipynb

README.md

dataset > aluguel.csv > data

```
1 Tipo;Bairro;Quartos;Vagas;Su  
2 Quitinete;Copacabana;1;0;0;40  
3 Casa;Jardim Botânico;2;0;1;10  
4 Conjunto Comercial/Sala;Barra  
5 Apartamento;Centro;1;0;0;15;8  
6 Apartamento;Higienópolis;1;0  
7 Apartamento;Vista Alegre;3;1  
8 Apartamento;Cachambi;2;0;0;50  
9 Casa de Condomínio;Barra da T
```

dataset >  aluguel.csv >  data

	Tipo	Bairro	Quartos	Vagas	Suites	Area	Valor	Condominio	IPTU
1	Quitinete	Copacabana	1	0	0	40	1700	500	60
2	Casa	Jardim Botânico	2	0	1	100	7000		
3	Conjunto Comercial/Sala	Barra da Tijuca	0	4	0	150	5200	4020	1111
4	Apartamento	Centro	1	0	0	15	800	390	20
5	Apartamento	Higienópolis	1	0	0	48	800	230	
6	Apartamento	Vista Alegre	3	1	0	70	1200		
7	Apartamento	Cachambi	2	0	0	50	1300	301	17
8	Casa de Condomínio	Barra da Tijuca	5	4	5	750	22000		
9	Casa de Condomínio	Ramos	2	2	0	65	1000		
10	Conjunto Comercial/Sala	Centro	0	3	0	695	35000	19193	3030
11	Apartamento	Centro	1	0	0	36	1200		
12	Apartamento	Grajaú	2	1	0	70	1500	642	74

Col 7: Valor

```
url = "dataset/aluguel.csv"
pd.read_csv(url)
```

[4]

✓ 0.1s

Python

...

Tipo;Bairro;Quartos;Vagas;Suites;Area;Valor;Condominio;IPTU

0

Quitinete;Copacabana;1;0;0;40;1700;500;60

1

Casa;Jardim Botânico;2;0;1;100;7000;;

2

Conjunto Comercial/Sala;Barra da Tijuca;0;4;0;...

3

Apartamento;Centro;1;0;0;15;800;390;20

4

Apartamento;Higienópolis;1;0;0;48;800;230;

...

...

1. Variável
2. Atribuição
3. Endereço relativo
4. pd – aliás do módulo pandas
5. Função read_csv()
6. Parâmetro da função read_csv(variável)

Qual o tipo da variável url?



```
print(type(url))
```

[5]



0.0s

```
... <class 'str'>
```

comma-separated values (csv)

As vezes não vem separado por vírgula, como nosso exemplo, veio separado por ponto e vírgula.

```
pd.read_csv(url, sep=";")
```

[6]

✓ 0.0s

Python

...

	Tipo	Bairro	Quartos	Vagas	Suites	Area	Valor	Condominio	IPTU
0	Quitinete	Copacabana	1	0	0	40	1700.0	500.0	60.0
1	Casa	Jardim Botânico	2	0	1	100	7000.0	NaN	NaN

```
pd.read_csv(P1urlX,P2sep=";"X)
```

Criando variáveis

```
dados = pd.read_csv(url, sep=";")
dados
```

[8]



0.1s

Python

...

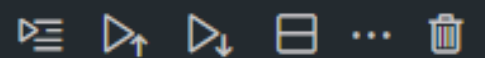
	Tipo	Bairro	Quartos	Vagas	Suites	Area	Valor	Condominio	IPTU
0	Quitinete	Copacabana	1	0	0	40	1700.0	500.0	60.0
1	Casa	Jardim Botânico	2	0	1	100	7000.0	NaN	NaN
2	Conjunto Comercial/Sala	Barra da Tijuca	0	4	0	150	5200.0	4020.0	1111.0
3	Apartamento	Centro	1	0	0	15	800.0	390.0	20.0
4	Apartamento	Higienópolis	1	0	0	48	800.0	230.0	NaN

Visualização não tá legal!

	Tipo	Bairro	Quartos	Vagas	Suites	Area	Valor	Condominio	IPTU
0	Quitinete	Copacabana	1	0	0	40	1700.0	500.0	60.0
1	Casa	Jardim Botânico	2	0	1	100	7000.0	NaN	NaN
2	Conjunto Comercial/Sala	Barra da Tijuca	0	4	0	150	5200.0	4020.0	1111.0
3	Apartamento	Centro	1	0	0	15	800.0	390.0	20.0
4	Apartamento	Higienópolis	1	0	0	48	800.0	230.0	NaN
...
32955	Quitinete	Centro	0	0	0	27	800.0	350.0	25.0
32956	Apartamento	Jacarepaguá	3	1	2	78	1800.0	800.0	40.0
32957	Apartamento	São Francisco Xavier	2	1	0	48	1400.0	509.0	37.0
32958	Apartamento	Leblon	2	0	0	70	3000.0	760.0	NaN

Visualizar apenas algumas linhas

Utiliza-se o método head (readi)



```
dados.head()
```

[9]



0.0s

Python

...

	Tipo	Bairro	Quartos	Vagas	Suites	Area	Valor	Condominio	IPTU
0	Quitinete	Copacabana	1	0	0	40	1700.0	500.0	60.0
1	Casa	Jardim Botânico	2	0	1	100	7000.0	NaN	NaN
2	Conjunto Comercial/Sala	Barra da Tijuca	0	4	0	150	5200.0	4020.0	1111.0
3	Apartamento	Centro	1	0	0	15	800.0	390.0	20.0
4	Apartamento	Higienópolis	1	0	0	48	800.0	230.0	NaN

A função `head()` retornar as cinco primeiras linhas. Caso necessite mostrar mais é só passar o número de linha desejada.

```
dados.head(15)
```

[10]



0.0s

Python

...

	Tipo	Bairro	Quartos	Vagas	Suites	Area	Valor	Condominio	IPTU
0	Quitinete	Copacabana	1	0	0	40	1700.0	500.0	60.0
1	Casa	Jardim Botânico	2	0	1	100	7000.0	NaN	NaN
2	Conjunto Comercial/Sala	Barra da Tijuca	0	4	0	150	5200.0	4020.0	1111.0
3	Apartamento	Centro	1	0	0	15	800.0	390.0	20.0
4	Apartamento	Higienópolis	1	0	0	48	800.0	230.0	NaN
5	Apartamento	Vista Alegre	3	1	0	70	1200.0	NaN	NaN
6	Apartamento	Cachambi	2	0	0	50	1300.0	301.0	17.0
7	Casa de Condomínio	Barra da Tijuca	5	4	5	750	22000.0	NaN	NaN
8	Casa de Condomínio	Ramos	2	2	0	65	1000.0	NaN	NaN
9	Conjunto Comercial/Sala	Centro	0	3	0	695	35000.0	19193.0	3030.0
10	Apartamento	Centro	1	0	0	36	1200.0	NaN	NaN
11	Apartamento	Grajaú	2	1	0	70	1500.0	642.0	74.0
12	Apartamento	Lins de Vasconcelos	3	1	1	90	1500.0	455.0	14.0
13	Apartamento	Copacabana	1	0	1	40	2000.0	561.0	50.0
14	Quitinete	Copacabana	1	0	0	27	1800.0	501.0	NaN

✓ Temos outro método, que mostra das últimas linhas

```
dados.tail()
```

[11]



0.0s

Python

...

	Tipo	Bairro	Quartos	Vagas	Suites	Area	Valor	Condominio	IPTU
32955	Quitinete	Centro	0	0	0	27	800.0	350.0	25.0
32956	Apartamento	Jacarepaguá	3	1	2	78	1800.0	800.0	40.0
32957	Apartamento	São Francisco Xavier	2	1	0	48	1400.0	509.0	37.0
32958	Apartamento	Leblon	2	0	0	70	3000.0	760.0	NaN
32959	Conjunto Comercial/Sala	Centro	0	0	0	250	6500.0	4206.0	1109.0

+ Code

+ Markdown



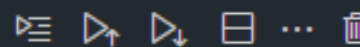
```
dados.tail(20)
```

[12] ✓ 0.0s

Python



	Tipo	Bairro	Quartos	Vagas	Suites	Area	Valor	Condominio	IPTU
32940	Apartamento	Barra da Tijuca	2	1	1	85	2250.0	1561.0	197.0
32941	Apartamento	Barra da Tijuca	3	2	0	98	2300.0	887.0	177.0
32942	Conjunto Comercial/Sala	Barra da Tijuca	0	1	0	55	2000.0	1096.0	NaN
32943	Apartamento	Barra da Tijuca	3	2	2	140	5500.0	1900.0	700.0
32944	Apartamento	Recreio dos Bandeirantes	3	2	1	88	1550.0	790.0	NaN
32945	Quitinete	Copacabana	1	0	0	22	1500.0	286.0	200.0
32946	Conjunto Comercial/Sala	Centro	0	0	0	140	4000.0	1412.0	496.0
32947	Apartamento	Leblon	3	0	1	80	3000.0	1010.0	249.0
32948	Conjunto Comercial/Sala	Centro	0	0	0	32	600.0	1035.0	83.0
32949	Apartamento	Ipanema	3	1	2	150	15000.0	1400.0	600.0
32950	Apartamento	Tijuca	1	0	0	28	1000.0	360.0	25.0
32951	Apartamento	Vila Valqueire	2	0	0	52	1000.0	550.0	NaN
32952	Casa de Condomínio	Barra da Tijuca	5	3	4	450	15000.0	1711.0	2332.0
32953	Apartamento	Méier	2	0	0	70	900.0	490.0	48.0
32954	Box/Garagem	Centro	0	0	0	755	14000.0	NaN	NaN
32955	Quitinete	Centro	0	0	0	27	800.0	350.0	25.0



```
"""
```

```
    Mostrando o tipo da minha variável
```

```
"""
```

```
type(dados)
```

[13] ✓ 0.0s

Python

... pandas.core.frame.DataFrame

Mova o post-it no Trello!

Trello

Áreas de trabalho

Recente

Marcado como favorito

Templates

Criar

Pesquisar

RP

Fundamentos Pandas

Quadro

Filtros

RP

Compartilhar

...

Informações do Projeto

O Problema

O DataSet

Adicionar um cartão

A Fazer

A FAZER

Data Science

1

Importar e conhecer a base de dados

Desenvolvimento

não há

Adicionar um cartão

Concluído

não há

Adicionar um cartão

Adicionar outra lista

INOVA TECH

Execução

Realização

SENAI

fapdf

Secretaria de Ciência, Tecnologia



tarefas

Ocultar itens marcados

Excluir

50%



importar os dados



explorar as características gerais dos dados

Importar e conhecer a base de dados



1/2

Até a próxima!

.... Exercício: pesquise sobre a palavra “pandas python no youtube”