

# Análise de dados com Python no dia a dia

Substituindo planilhas por código

Rodrigo Amaral

DCOMP/UFS

23 de agosto de 2019

São Cristóvão - Sergipe



# Quem sou eu

## Rodrigo Amaral

- Analista de Tecnologia da Informação no TRT da 20ª Região (SE)
- Mestre em Ciência da Computação pela UFS
- Coorganizador do Python User Group Sergipe (PUG-SE)
- Fã de música, livros e basquete 🎸 📚 🏀



rodrigo@rodrigoamaral.net



amaral101



amaral101

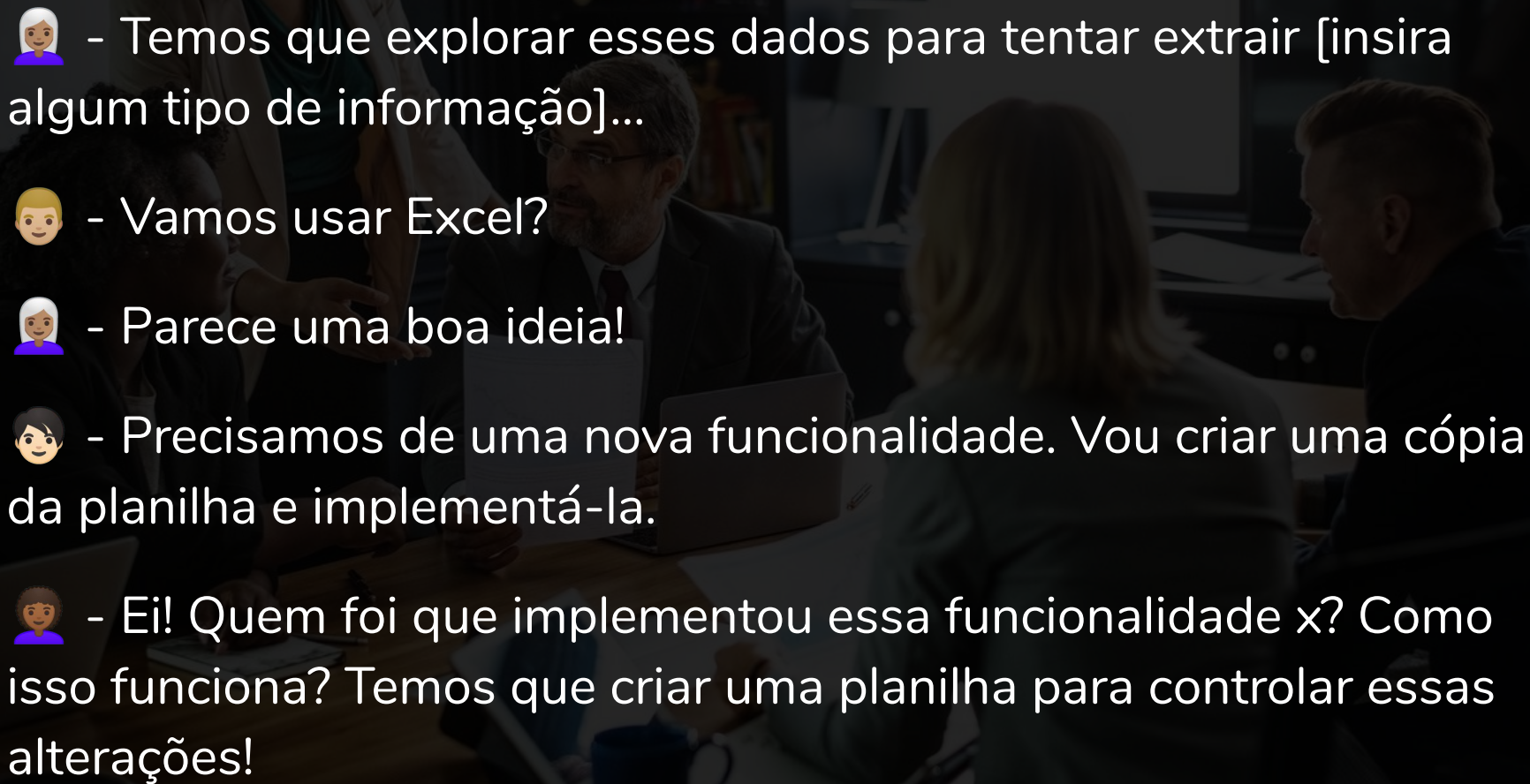







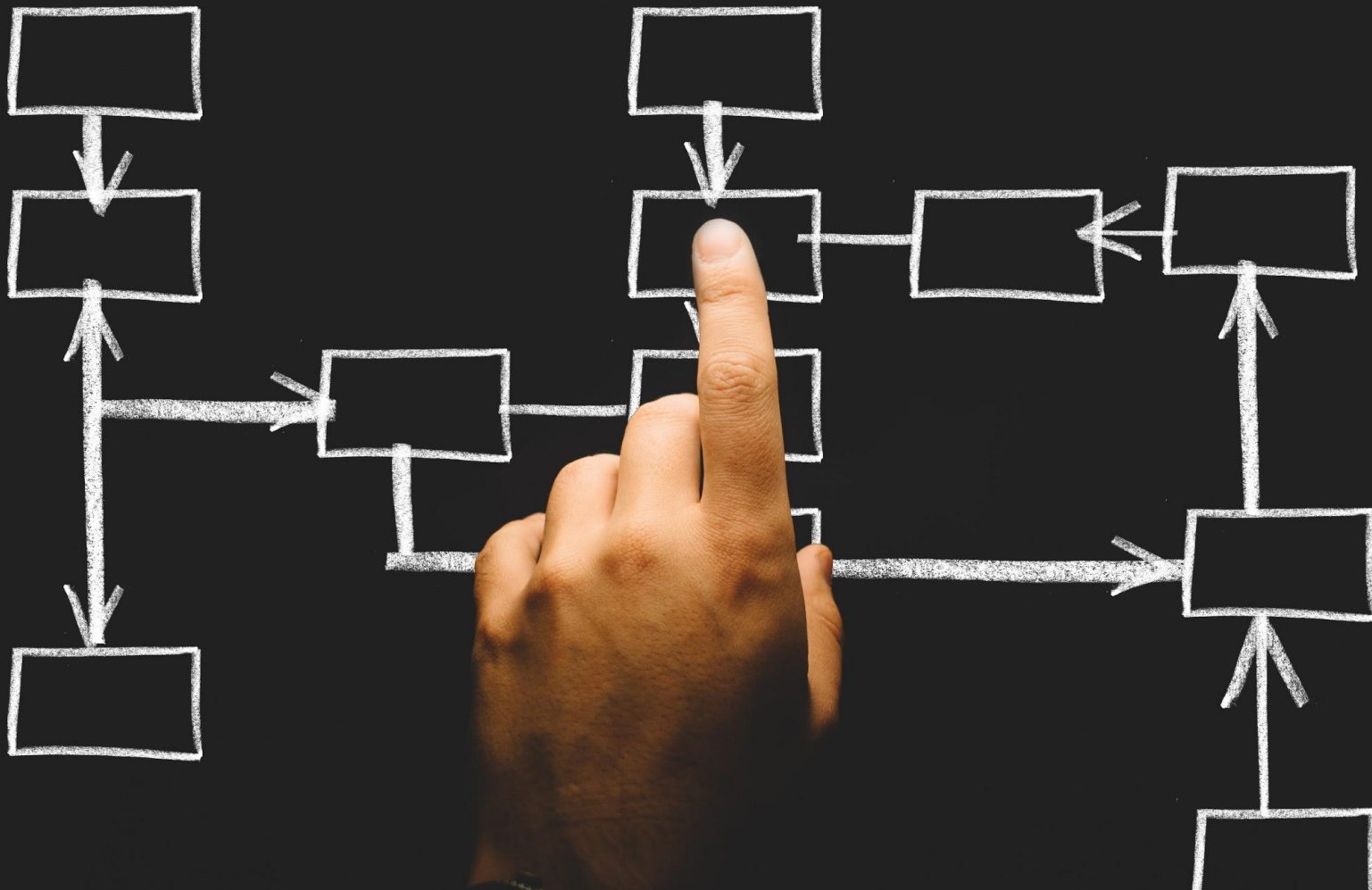
rodrigoamaral



rodrigoamaral

# Um dia comum no escritório...

- 
- A group of four office workers are gathered around a table in a dimly lit office. A woman with curly hair is standing and pointing at a laptop screen. A man with glasses is sitting at the table, looking at the screen. A woman with blonde hair is sitting with her back to the camera, looking at the screen. A man with short hair is sitting on the right, looking at the screen. There are papers, a blue mug, and a laptop on the table.
-  - Temos que explorar esses dados para tentar extrair [insira algum tipo de informação]...
  -  - Vamos usar Excel?
  -  - Parece uma boa ideia!
  -  - Precisamos de uma nova funcionalidade. Vou criar uma cópia da planilha e implementá-la.
  -  - Ei! Quem foi que implementou essa funcionalidade x? Como isso funciona? Temos que criar uma planilha para controlar essas alterações!



















1

# Existe algum fluxo nas operações da planilha?

# Documents library

TPS

Arrange by: P

Name	Date modified	Type	Size
 Final-TPS-Report-A.xlsx	3/15/2016 12:40 PM	Microsoft Excel ...	567 KB
 Report-TPS-VPv2.xlsx	2/26/2016 11:33 A...	Microsoft Excel ...	757 KB
 TPS_Rpt-v4.xlsx	6/17/2015 5:03 PM	Microsoft Excel ...	109 KB
 TPS-Report-2-3-2016.xlsx	2/3/2016 4:09 PM	Microsoft Excel ...	726 KB
 TPS-Report-5-2015v1.xlsx	5/29/2015 1:50 PM	Microsoft Excel ...	611 KB
 TPS-Report-Final-Final.xlsx	12/10/2015 3:40 PM	Microsoft Excel ...	541 KB
 TPS-Report-Final-Finalv2.xlsx	12/9/2015 9:23 PM	Microsoft Excel ...	73 KB
 TPS-Report-Jun-2015v2.xlsx	7/17/2015 3:52 PM	Microsoft Excel ...	53 KB
 TPS-Report-revc.xlsx	7/17/2015 4:08 PM	Microsoft Excel ...	50 KB
 TPS-Report-v3.1.xlsx	5/27/2015 10:53 A...	Microsoft Excel ...	283 KB
 TPS-Report-v3.xlsx	5/31/2015 8:24 PM	Microsoft Excel ...	875 KB
 TPS-report-v8.xlsx	7/15/2015 7:23 PM	Microsoft Excel ...	46 KB
			KB
			KB
			KB
 TPS-VII-Dec.xlsx	12/2/2015 3:33 PM	Microsoft Excel ...	KB

2 Qual a versão que está valendo?





3

Como a planilha funciona? De onde vieram os dados?





4

O conjunto de dados tem MILHÕES de registros. Como fazer?

# Copiar e colar não é uma opção

Situações que imploram por automatização de planilhas:

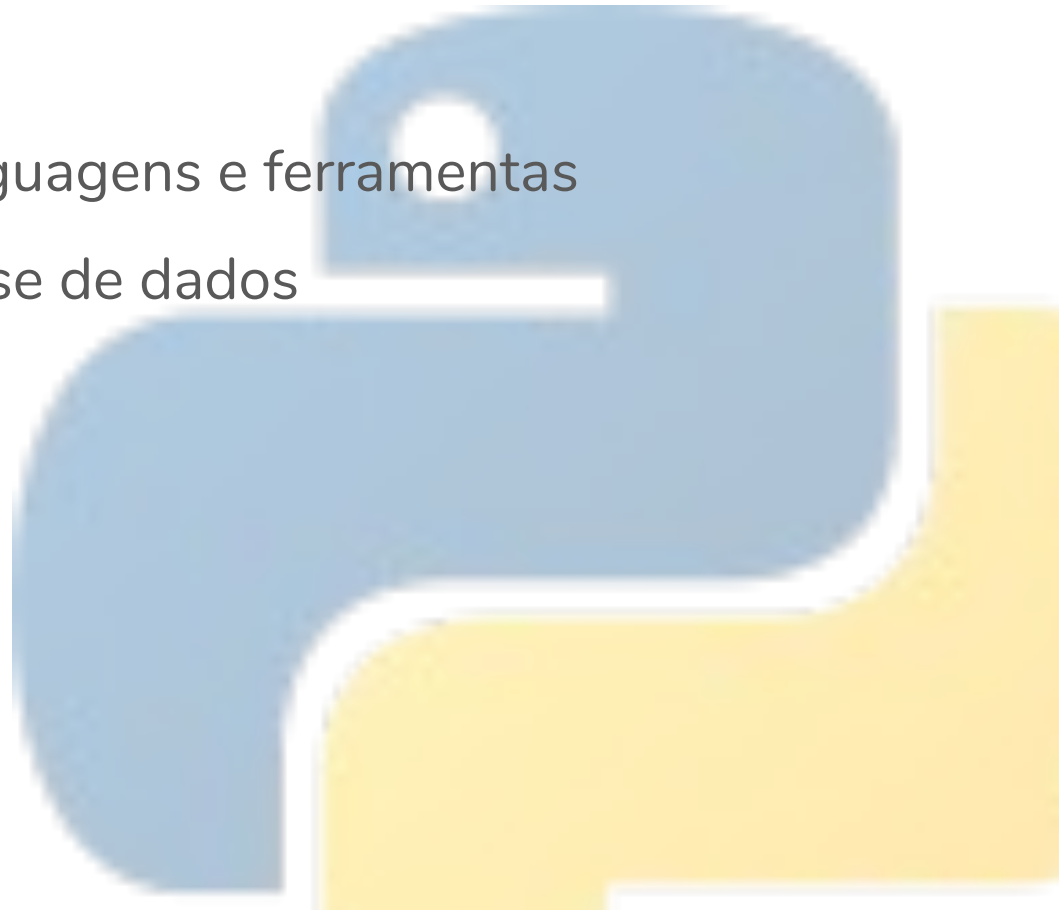
- O tempo disponível é curto
- O processo é tedioso e sujeito a erros
- Os conjuntos de dados excedem a capacidade da planilha
- A tarefa envolve muita manipulação de texto





# Por que Python?

- Legível e expressiva
- Fácil de aprender
- Múltiplos propósitos
- Integração com outras linguagens e ferramentas
- Ampla utilização em análise de dados



# “Estudos comprovam!”

## What do you use Python for?



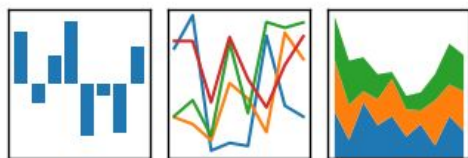
JetBrains Developer Survey 2019

# Ferramentas para ciência de dados

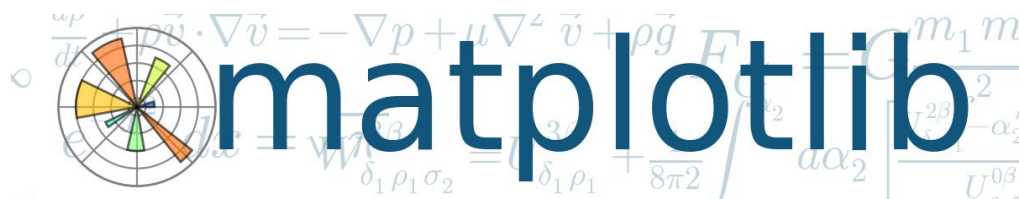


pandas

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



IP[y]:  
IPython



seaborn

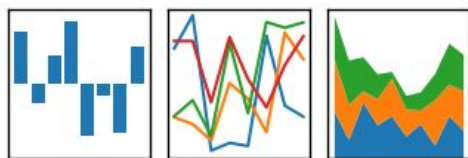


# Ferramentas para ciência de dados

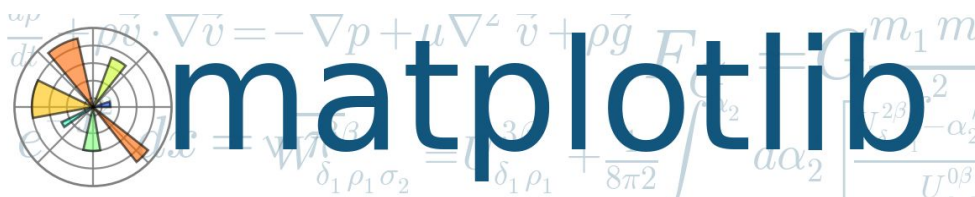
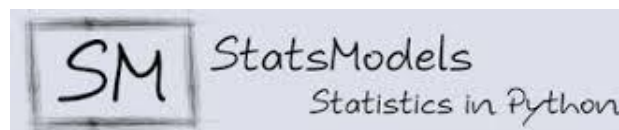


pandas

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



IP[y]:  
IPython



seaborn

**Vamos à prática!**

# Boas práticas

- Use um sistema de controle de versões (e.g. git) para o código produzido
- Mantenha código e dados separados
- Escreva – e mantenha atualizada – documentação para os trechos de código mais complexos



# Mas calma... também há alguns desafios

- Jupyter
  - Células executadas fora de ordem
- Aprender uma nova tecnologia
  - Como convencer as pessoas de mais essa tarefa
- Formatação
  - Ficamos sem os recursos de formatação de texto
- Distribuição
  - Como compartilhar a solução com outras pessoas de maneira fácil?





O PUG-SE é uma iniciativa comunitária que tem o objetivo de reunir os desenvolvedores e demais interessados na linguagem de programação Python e em suas tecnologias associadas.

<http://se.python.org.br>



[@pugse](https://t.me/pugse)

**<http://rodrigoamaral.net/palestras/>**



# Obrigado!

Dúvidas?

Perguntas?

# Referências

1. MOFFITT, Chris. “Escaping the Excell Hell with Python and Pandas”  
<<https://github.com/chris1610/pbpython/blob/master/presentations/Escaping-Excel-Hell-with-Python-and-Pandas.pdf>>
2. JetBrains. “The State of Developer Ecosystem 2019”.  
<<https://www.jetbrains.com/lp/devecosystem-2019/>>
3. pandas: powerful Python data analysis toolkit <<https://pandas.pydata.org/pandas-docs/stable/>>
4. GRUS, Joel. “I Don’t Like Notebooks”. Slides  
<<https://docs.google.com/presentation/d/1n2RIMdmv1p25Xy5thJUhkKGvjtV-dkAIsUXP-AL4ffl/edit?usp=sharing>>, Vídeo <<https://www.youtube.com/watch?v=7jiPeIFXb6U>>