Python Data Analysis Mastery

Prof. Nicksson Freitas

Stack

Aula 03. Python para Análise de Dados

Prof. Nicksson Freitas

Stack

Objetivos

- Conhecer as soluções, ferramentas e linguagem de programação para construir aplicações envolvendo dados ou trabalhar com ciência de dados
- Conhecer a linguagem Python com o foco na análise de dados

Resumo

1. Ferramentas para Ciência de Dados
2. Linguagem de Programação
3. Por que Python?
4. Kit de ferramentas do Python
5. Vamos praticar?

1. Ferramentas para Ciência de Dados?

Stack

Ferramentas para Ciência de Dados

Ferramentas para Ciência de Dados

1st International Conference on Electrical and Information Technologies ICEIT'2015

Open Source Data Mining Tools

A Comparative Study

Hussah A. Al-Odan, Ahmad A. Al-Daraiseh King Saud
University Riyadh, Saudi Arabia
halodan@ksu.edu.sa, creepymaster@yahoo.com

Saiba mais: https://ieeexplore.ieee.org/abstract/document/7162956

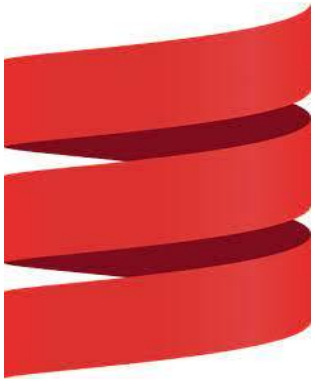
		WEKA	Orange	RapidMiner	RStudio	KNIME
Platform Support	Windows	√	√	√	√	√
	Mac	√	√	√	√	√
	Linux	√	√	√	√	√
Interface	Text	-	-	-	-	-
	GUI	-	-	-	√	-
	Interactive GUI	√	√	√	-	√
Installation Process	Single Package	√	√	-	-	√
	Multi Package	-	-	√	√	-
	Online / Server Application	-	-	√	√	-
	Flat Files	-	-	-	-	-
	Developer Version	√	√	√	√	√
Data Sources	MS Access	√	-	√	√	√
	MS Excel	-	-	√	-	√
	MySQL	√	√	-	√	√
	ARFF	√	√	√	√	√
	CSV	√	√	√	√	√
Supported Algorithms	Decision Trees	√	√	√	√	√
	Linear / Statistical	√	√	√	√	√
	Bayes	√	√	√	√	√
	Neural Networks	√	√	√	√	√
	Association Rules	√	√	√	√	√
	K Means	√	√	√	√	√
	Nearest Neighbor	√	√	√	√	√
Output	Bar Charts	√	√	√	√	√
	Pie Charts	√	√	√	√	√
	Scatter Plots	√	√	√	√	√
	Classification Trees	√	√	√	√	√

(AL-ODAN; AL-DARAISEH, 2015)

2. Linguagem de Programação

Stack

Linguagem de Programação para DS

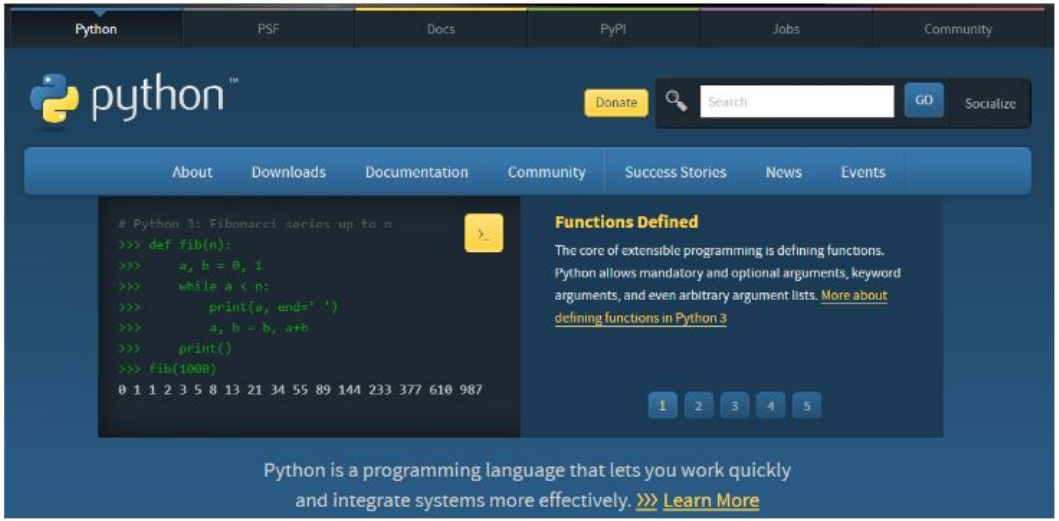


3. Por que Python?

Por que Python?

Quais São as Linguagens de Programação Mais Usadas em 2023?

1. Python



Saiba Mais

Dados de Vagas do Linkedin

Requisitos e Qualificações:

- Conhecimentos avançados em Probabilidade e Estatística;
- Experiência avançado em Linguagem de Programação (Python, SQL);
- Experiência com processos ETL (Extract, transform, load) e engenharia de features;
- Experiência consistente em desenvolvimento de modelos e pipelines de Machine Learning com Sklearn e PySpark;
- Conhecimento avançado em arquitetura de software ou serviço de dados (batch, online, real-time, etc.);
- Experiência trabalhando em uma base de código com colaboradores e com um sistema de versionamento (ex. GitHub);
- Experiência com testes A/B e experimentação em geral;
- Perfil crítico, analítico e data-driven;
- Conhecimento em desenvolvimento e boas práticas de software (reutilizável, testável e documentado).

O Que Esperamos De Você

- Extração de dados em bancos de dados tabulares (SQL, BigQuery);
- Manipulação e tratamento de bases massivas de dados (SQL, Python, Pandas, PySpark);
- Conhecimento em desenvolvimento utilizando a linguagem Python, além de familiaridade com bibliotecas para manipulação, análise e visualização de dados (Pandas, NumPy, Matplotlib, Seaborn, Plotly);
- Desenvolvimento de modelos preditivos de classificação, regressão e clusterização, utilizando técnicas de machine learning (scikit-learn, Tensorflow);
- Ter atuado com prévia em análise e modelagem estatística;
- Cabeça de dono;
- Trabalho em equipe;
- Boa comunicação e escuta ativa;
- Perfil analítico.

Job Requirements:

- Bachelor's/Master's degree in Engineering, Computer Science (or equivalent experience)
- At least 3+ years of relevant experience as a Data Scientist
- Expertise in working with languages like Python and R
- Thorough knowledge of quantitative data analysis. report writing and presenting findings
- Deep understanding of data querying languages like SQL
- Utilize analytical skills to collect, organize, analyze, and disseminate large amounts of data with accuracy
- Knowledge of updated applied statistics or experimentation methods
- Capacity to handle stats-based modeling and ML modeling

About The Role

- Conduct independent reviews of quantitative models based on statistical, machine learning and AI techniques, with stronger focus on credit models;
- Provide effective challenge, identify risks, enhancement opportunities, and engage with other Data Scientists and Business Analysts to strengthen our decision making tools;
- Develop playbooks and toolkits (Python, Scala, SQL, etc) to optmize model reviews, ongoing models monitoring, and assess the impact of models in decisions;
- Contribute to the consolidation of Nubank's Model Risk Management and Model Review frameworks with autonomy and creativity;
- Discuss and report model risk status and independent opinions on models with different stakeholders, including senior managers and regulators;
- Ensure the team maintains a high level of technical excellence.

Stack

Empresas que usam Python

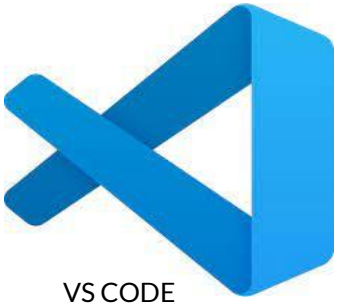
4. Kit de Ferramentas para Desenvolvimento usando Python

Stack

Quais são as ferramentas principais?

- Interpretador Python
  - Ferramenta para desenvolver projetos usando a linguagem python
- Gerenciador de pacotes e ambientes
  - Gerencia os ambientes virtuais e pacotes que serão utilizados no desenvolvimento de um projeto
- Controle de versionamento
  - Gerenciar as alterações do código ao longo do tempo

Interpretadores Python e IDEs

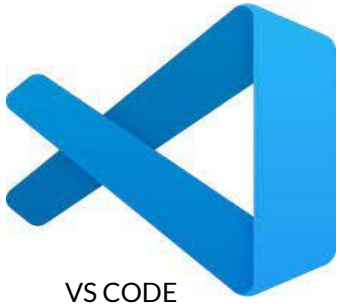


VS CODE

O que eu uso no dia a dia?

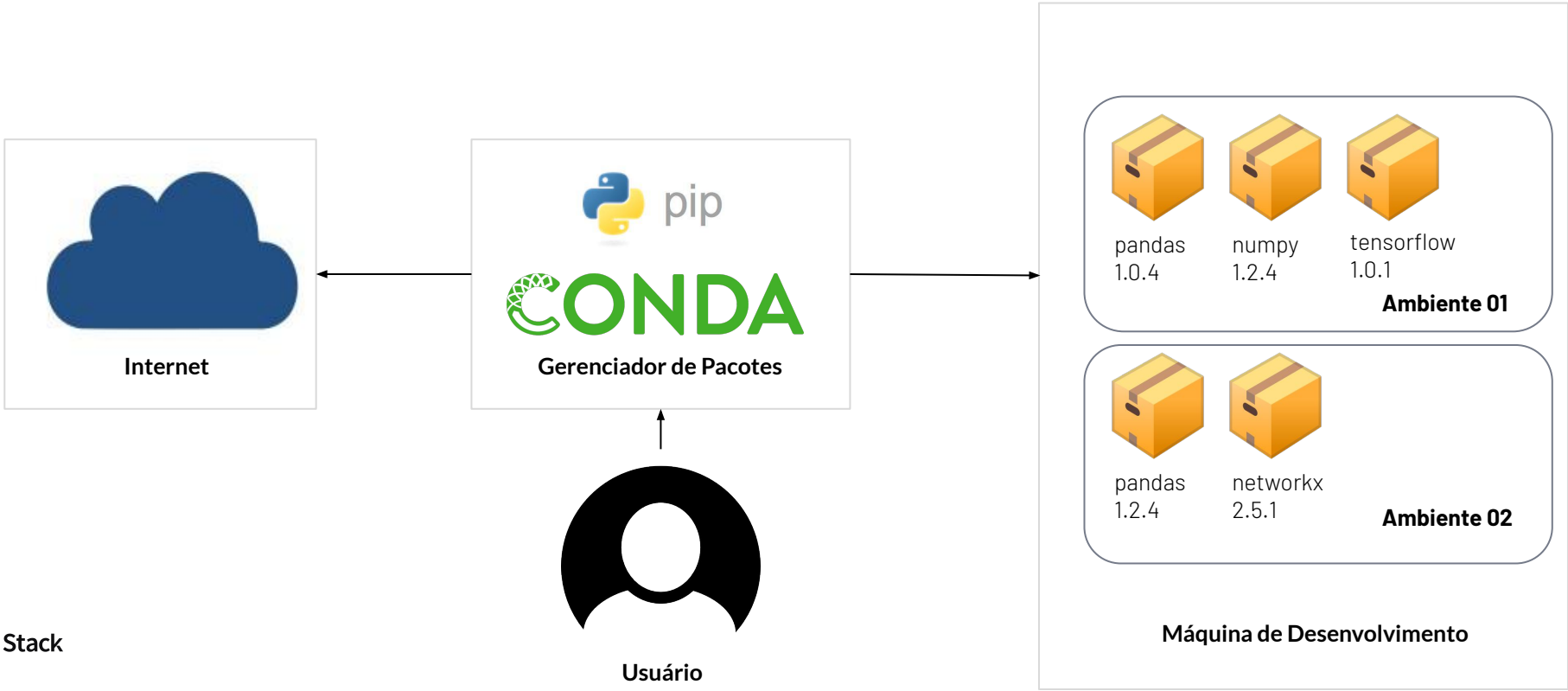
IPython

jupyter lab



VS CODE

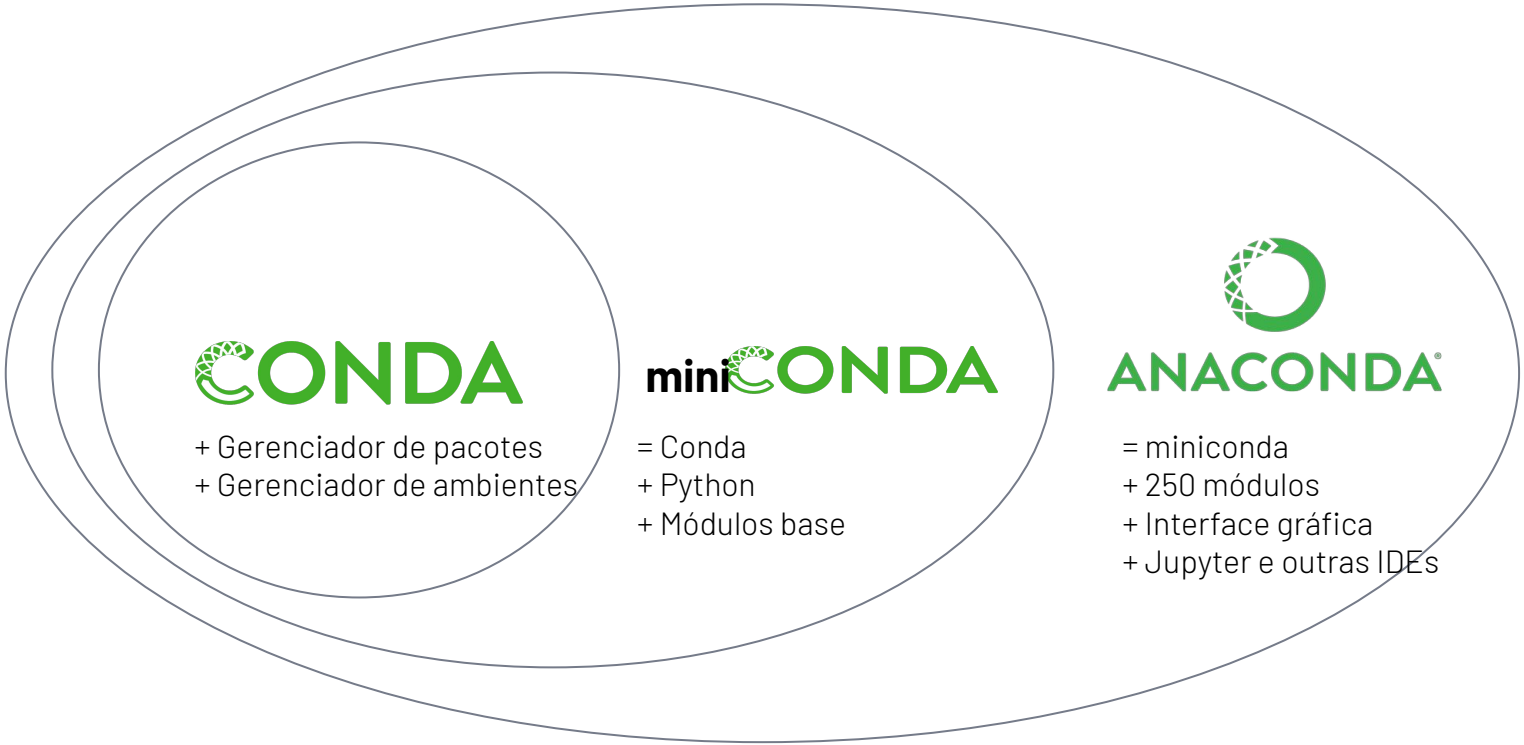
Para que serve um Gerenciador de Pacotes



Gerenciador de Pacotes e Ambientes

Conda	Pip
Gerenciador de pacotes	Gerenciador de pacotes
Gerenciador de ambientes	Depende de instalação externa (pipenv)
Suporta pacotes de outras linguagem	Suporta apenas pacotes Python
Realiza checagem de dependência	Não faz checagem de dependência
+2 mil pacotes no Anaconda Cloud	+300 mil pacotes no PyPi

Ecossistema conda



Controle de versionamento

6. Stack de Python para Data Science

Principais bibliotecas

- Análise estatística, manipulação de dados e modelagem de dados
  - Numpy
  - Pandas
  - Scipy
  - Statsmodel
- Visualização de dados
  - Matplotlib
  - Seaborn
  - Plotly

6. Vamos na prática?

Stack

Vamos preparar seu ambiente?

1. Instale um interpretador Python
    a. Anaconda
        i. https://www.anaconda.com/
    b. Google colab
        i. https://colab.research.google.com/

2. Baixe os dados do curso:
    c. https://github.com/ferramentas-stackacademy/Python-Data-Analysis-Mastery

Considerações Finais

Stack

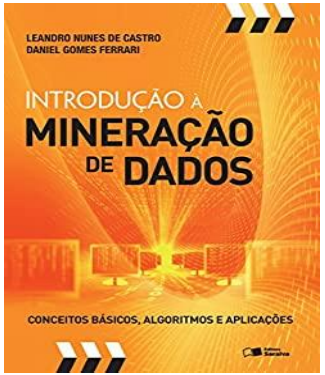
Considerações finais

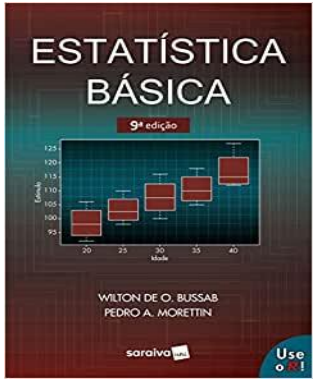
- Você conheceu diferentes ferramentas para trabalhar com Ciência de Dados
- Você conheceu as principais linguagem para trabalhar com dados
- Você conheceu o kit de ferramentas para trabalhar com Ciência de Dados usando Python e as principais bibliotecas para análise e visualização de dados

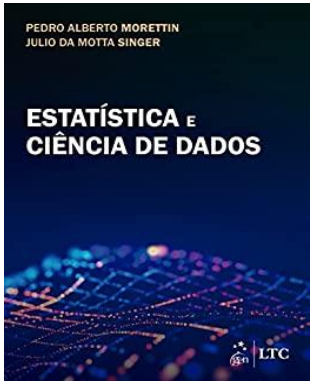
Referências

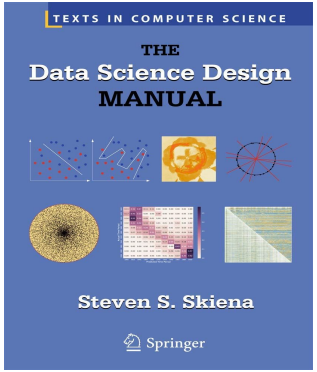
Stack

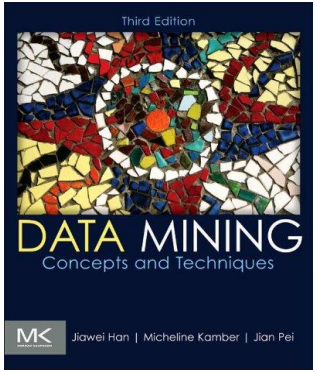
Bibliografia fundamental

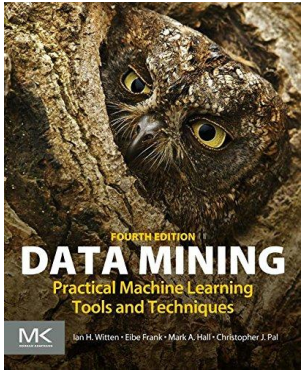






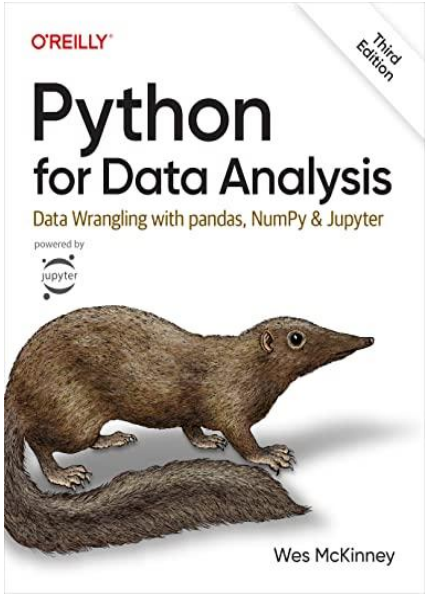


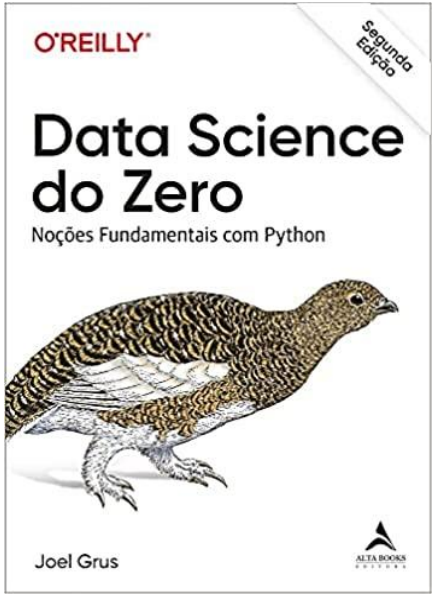


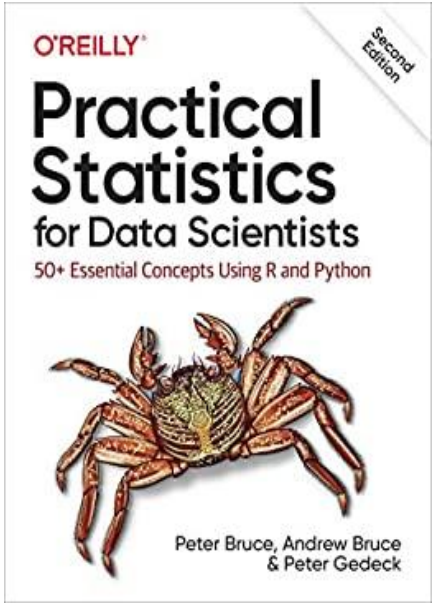


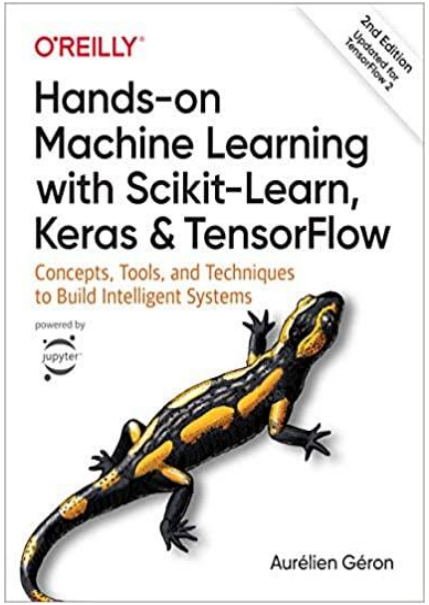
Stack

Bibliografia técnica









OBRIGADO