Introdução ao Anaconda, Jupyter Notebook e Python

Do conceito à instalação e manipulação

Quem Sou Eu

Uma breve descrição

APRESENTAÇÃO

Mateus Rocha

Técnico em Informática

IFPB - 2019

Estatístico

UEPB - 2024

Cientista de Dados e Professor

ASN-Rocks - Atualmente

Ambiente de Desenvolvimento

Dos editores de texto até o Anaconda

ONDE ESCREVER CÓDIGO?

Programar não é apenas escrever texto; é estabelecer uma comunicação entre a sua ideia e o processador do computador. **O ambiente de desenvolvimento serve como o tradutor e o espaço de trabalho nessa conversa.**

O Código é a Receita: Você pode escrevê-lo em qualquer lugar (até num guardanapo/Bloco de Notas).

O Ambiente é a Cozinha: Ele oferece o fogão (o motor que executa o código), os utensílios (ferramentas de correção) e a bancada organizada (onde você visualiza os arquivos).

IDE (Integrated Development Environment)

As IDEs são ambientes completos, como uma "oficina mecânica" cheia de ferramentas avançadas. Elas não servem apenas para escrever texto, mas oferecem tudo o que é necessário para grandes projetos.

Características

Possuem preenchimento automático de código (IntelliSense), ferramentas de depuração (debug) para encontrar erros, e integração com bancos de dados.

Exemplos

PyCharm (específica para Python), IntelliJ IDEA (muito usada para Java) e Visual Studio (completo para C# e C++).

EDITORES DE CÓDIGO

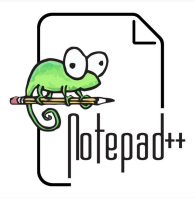
São muito mais rápidos e leves que uma IDE, mas dependem da instalação de extensões (plugins) para ganhar poderes.

Características

Focam na velocidade e na versatilidade. Você pode usar o mesmo editor para Python, HTML, CSS ou JavaScript apenas mudando a configuração.

Exemplos

VS Code (Visual Studio Code) — o mais popular atualmente — e o Sublime Text, notepad ++.



JUPYTER NOTEBOOK

O Jupyter é um ambiente de desenvolvimento interativo que funciona no navegador. Ele rompe com o modelo tradicional de "escrever um arquivo de texto longo e rodar tudo de uma vez", **permitindo que você execute o código em pedaços (células).**

Características

Permitem a mistura de código vivo, equações, textos explicativos e visualizações ricas (gráficos e tabelas). Sua grande vantagem é a execução modular, onde você não precisa rodar o projeto inteiro para testar apenas uma pequena alteração.

Exemplos

Jupyter Notebook (local, via Anaconda), Google Colab (versão em nuvem do Google) e Kaggle Kernels.

O QUE É O ANACONDA?

O Anaconda é a principal distribuição científica da linguagem Python. Ele organiza onde cada ferramenta fica, garante que elas tenham o que precisam para funcionar e evita conflitos entre elas.

Plataforma "Tudo-em-Um"

Ao instalar o Anaconda, você já recebe o Python, o Jupyter Notebook, o Spyder e as principais bibliotecas de ciência de dados pré-configuradas.

Gestão de Ambientes (Conda)

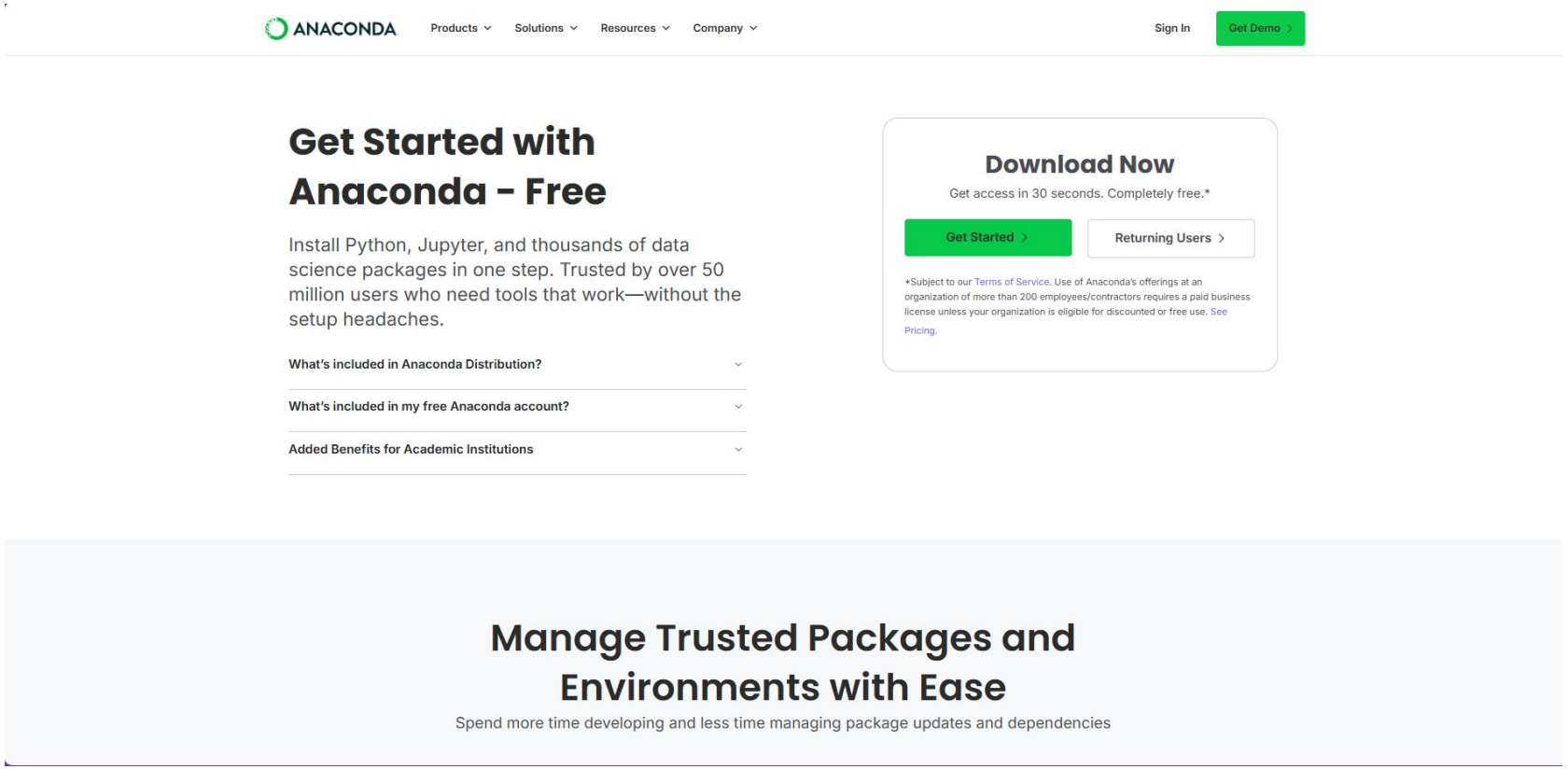
Permite criar "bolhas" isoladas para cada projeto. Você pode ter um projeto de Machine Learning usando uma versão X de uma biblioteca e outro projeto usando a versão Y, sem que um interfira no outro.

Baixando as Ferramentas

Entendendo como manipular o Anaconda e o Jupyter Notebook

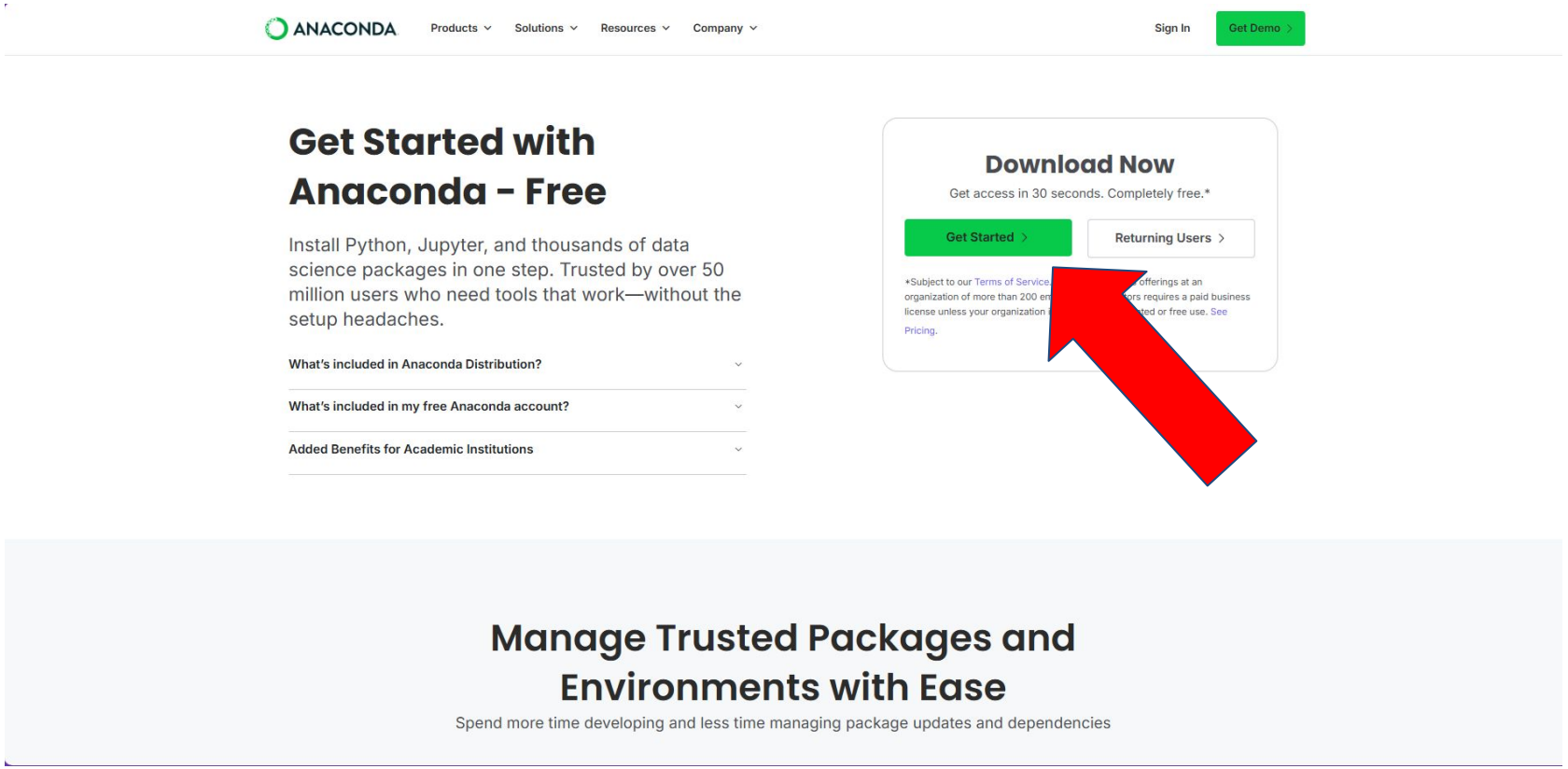
COMO BAIXAR?

Primeiramente Acesse o site: https://www.anaconda.com/download



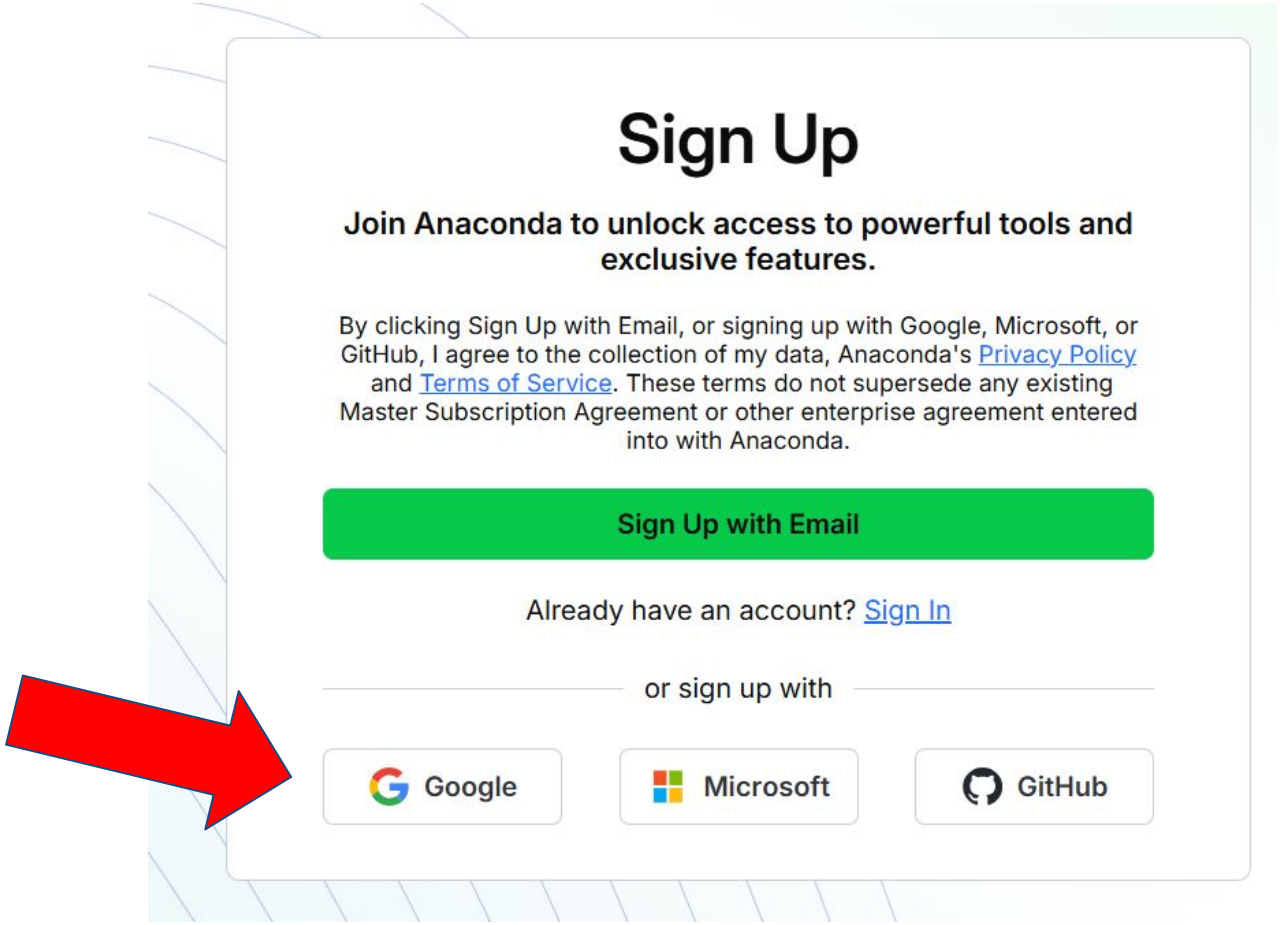
COMO BAIXAR?

Clique em **Get Started:**



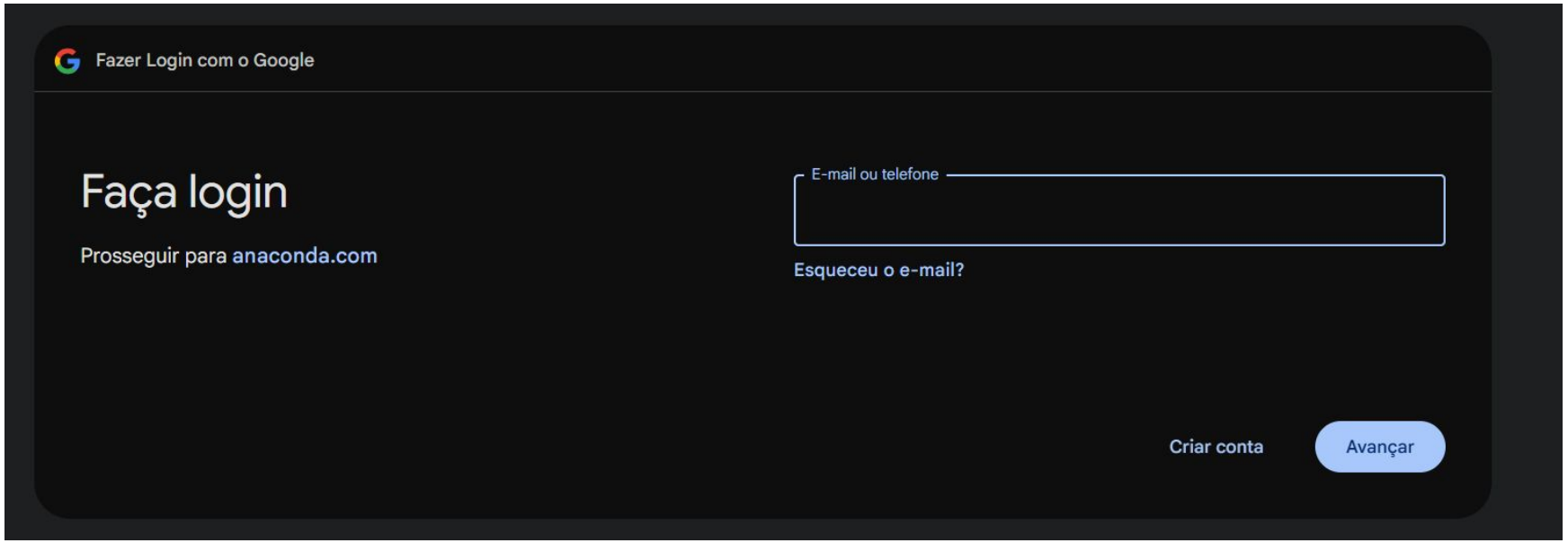
COMO BAIXAR?

Você precisará de uma conta para baixar o ANACONDA, vamos clicar no **Google por exemplo**



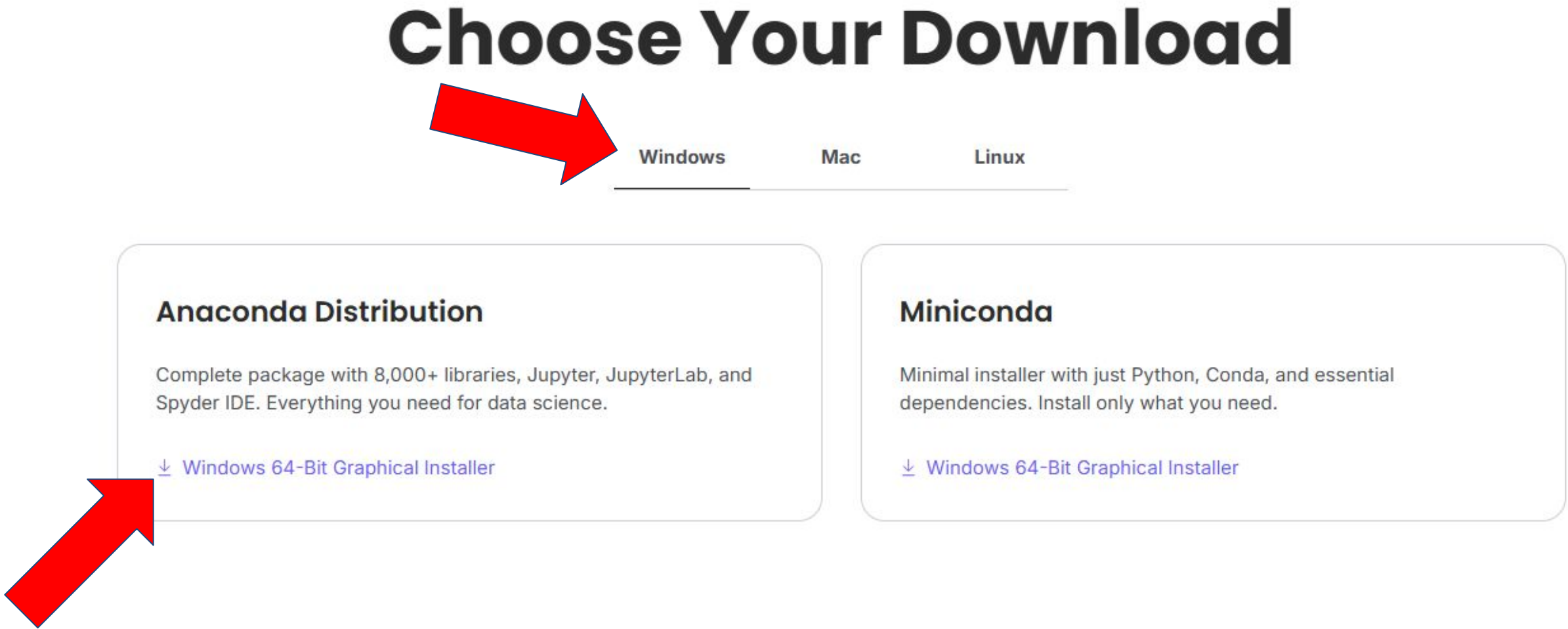
COMO BAIXAR?

Faça o login com a sua conta **Google**:



COMO BAIXAR?

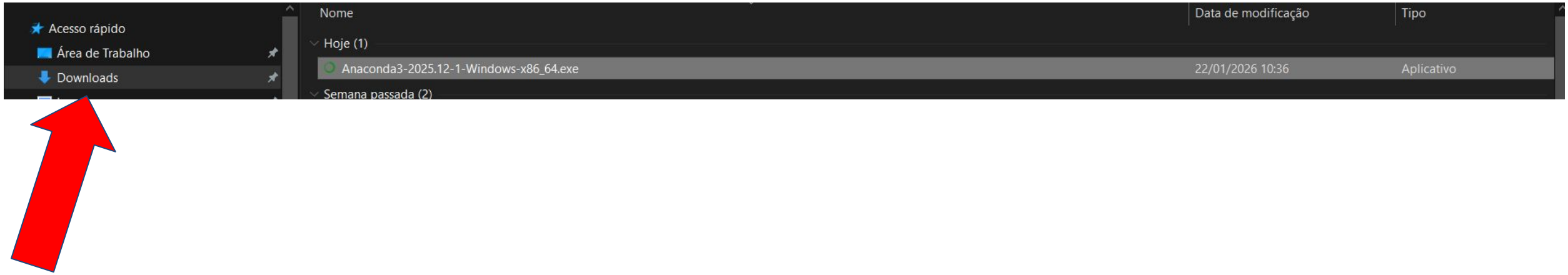
Primeiro escolha o sistema operacional e depois baixe o **Anaconda Distribution**



COMO BAIXAR?

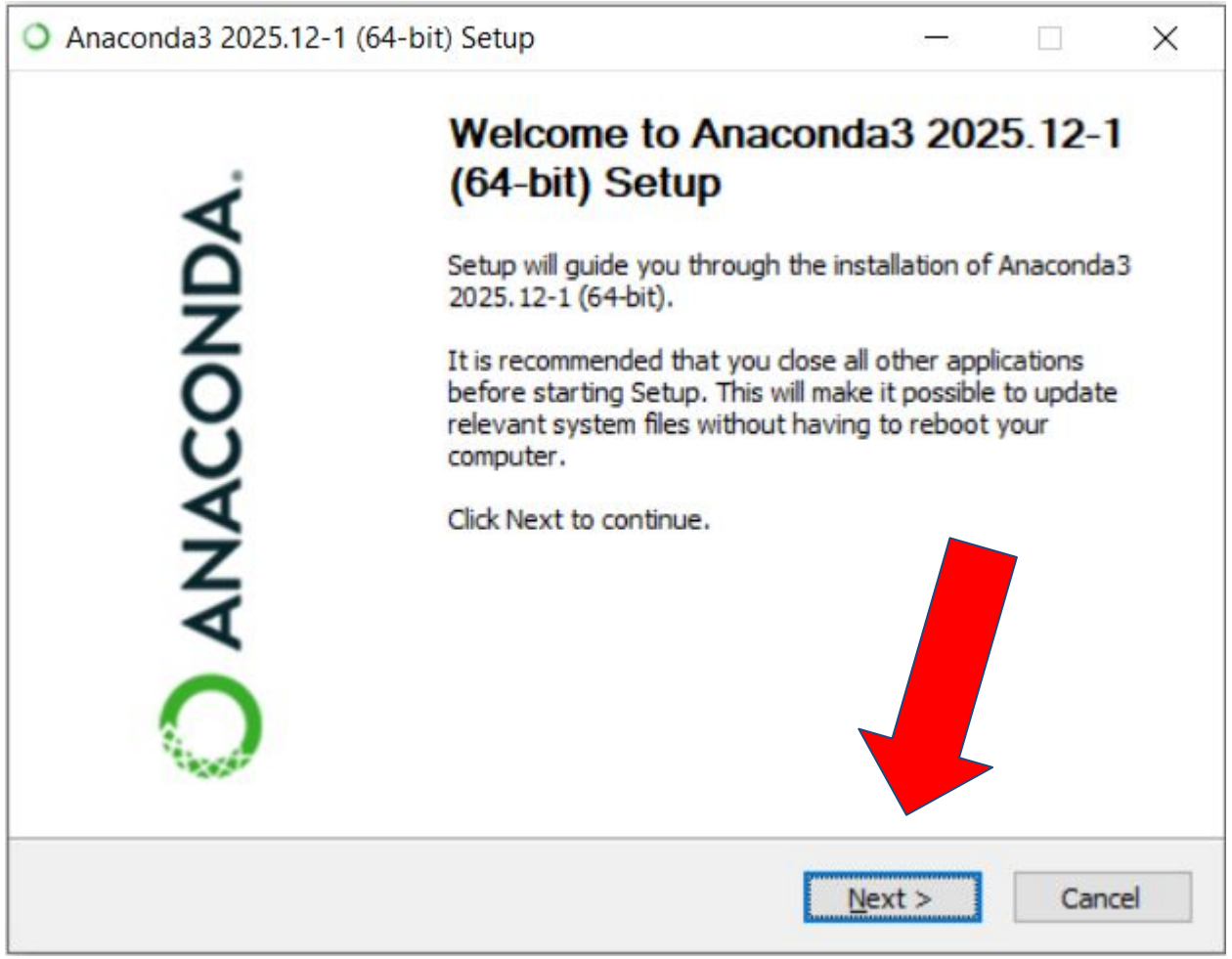
Vá na aba de Download do seu computador e clique no instalador

Anaconda3-2025.12-1-Windows-x86_64.exe



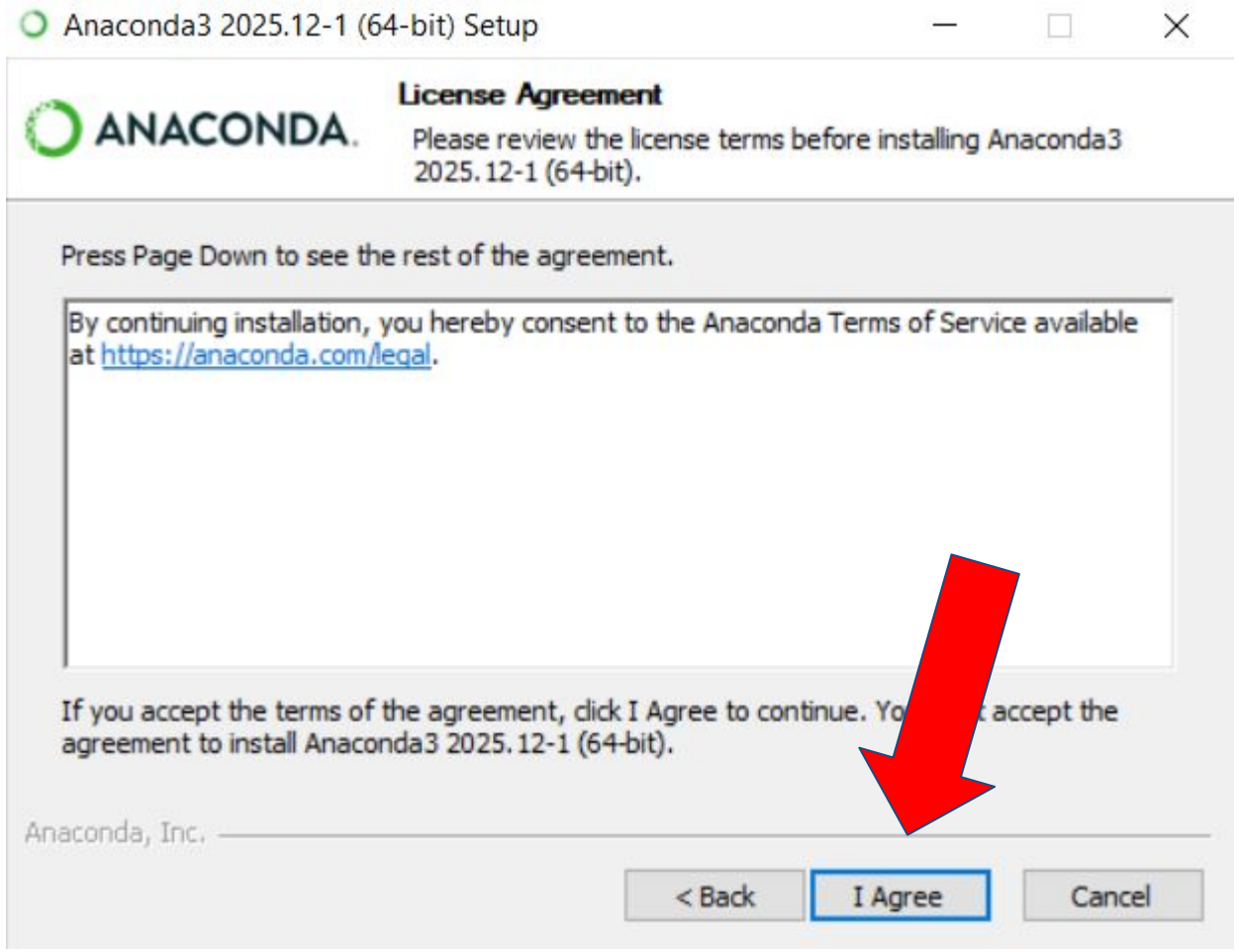
COMO BAIXAR?

Clique em **next**



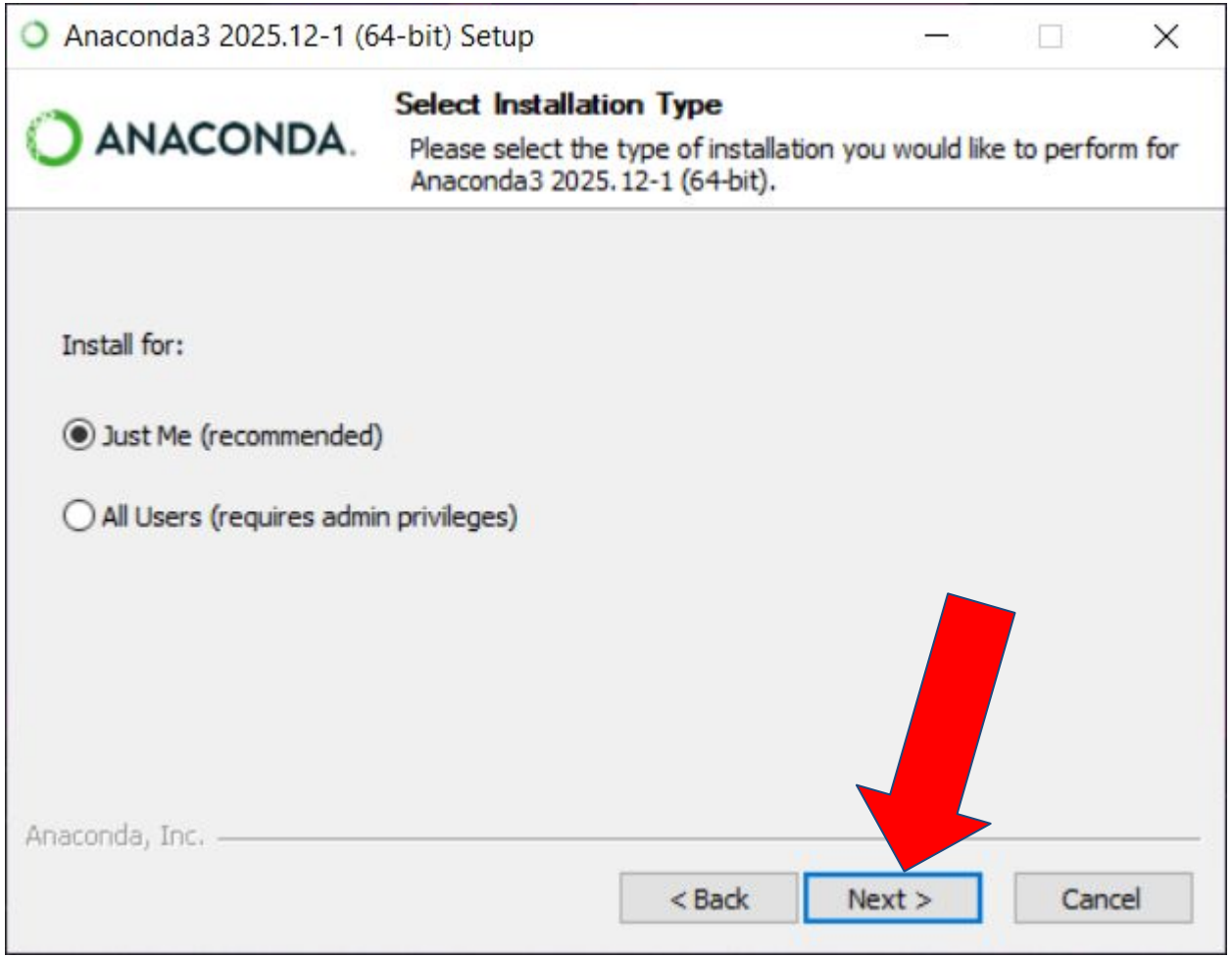
COMO BAIXAR?

Clique em **I Agree**



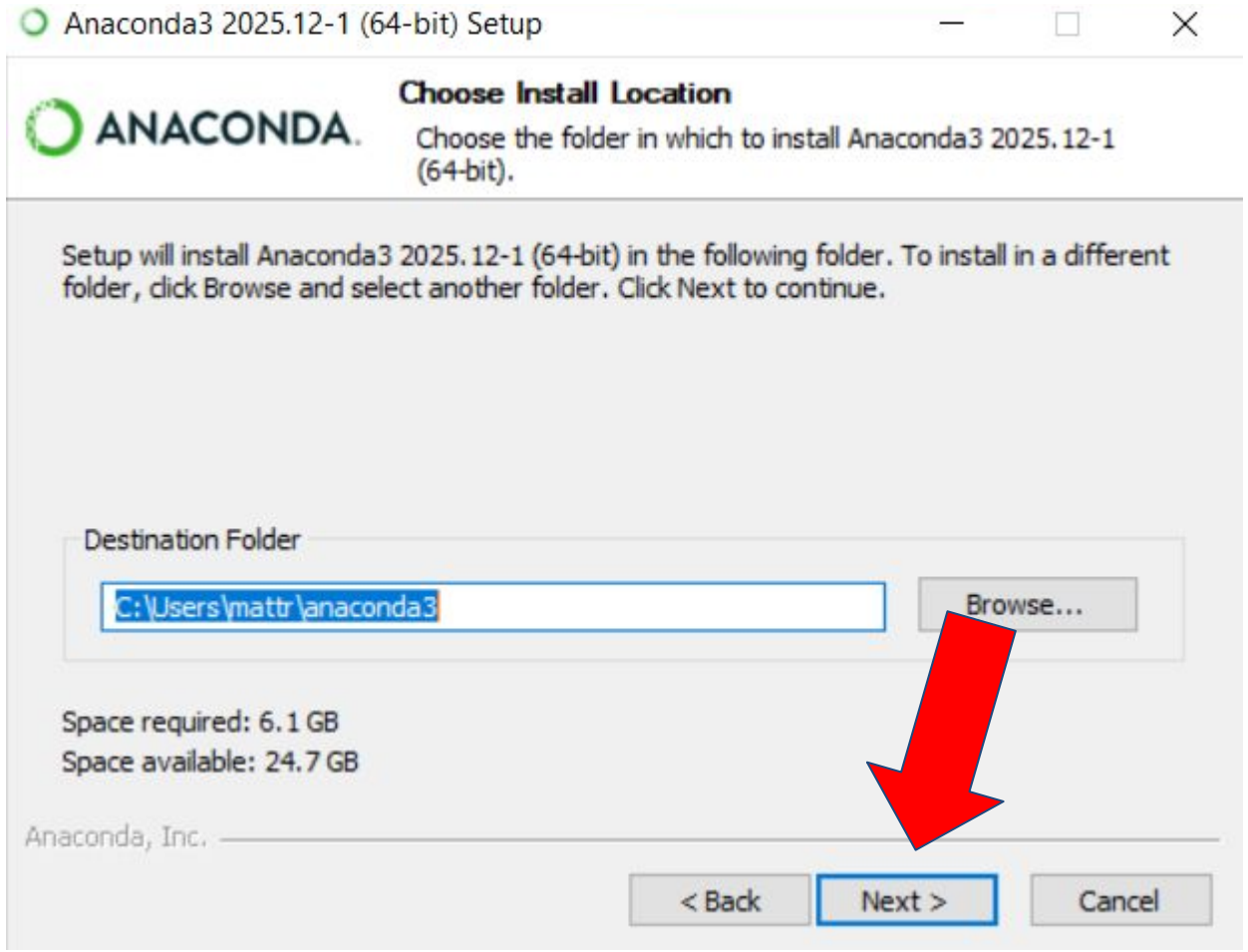
COMO BAIXAR?

Deixe a opção **Just Me** e clique em **Next**



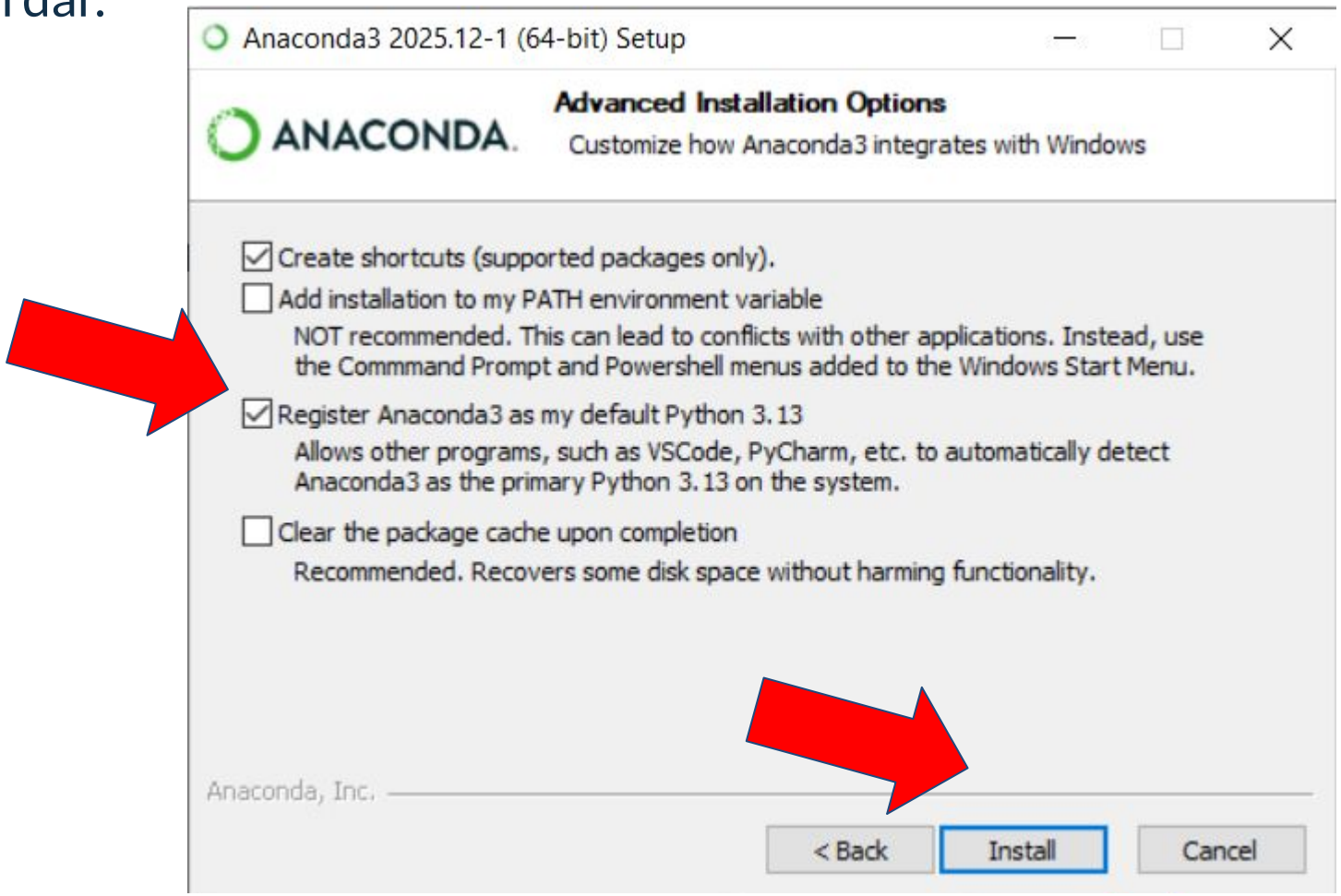
COMO BAIXAR?

Deixe a opção padrão do diretório de destino onde o Anaconda ficará e clique em **Next**



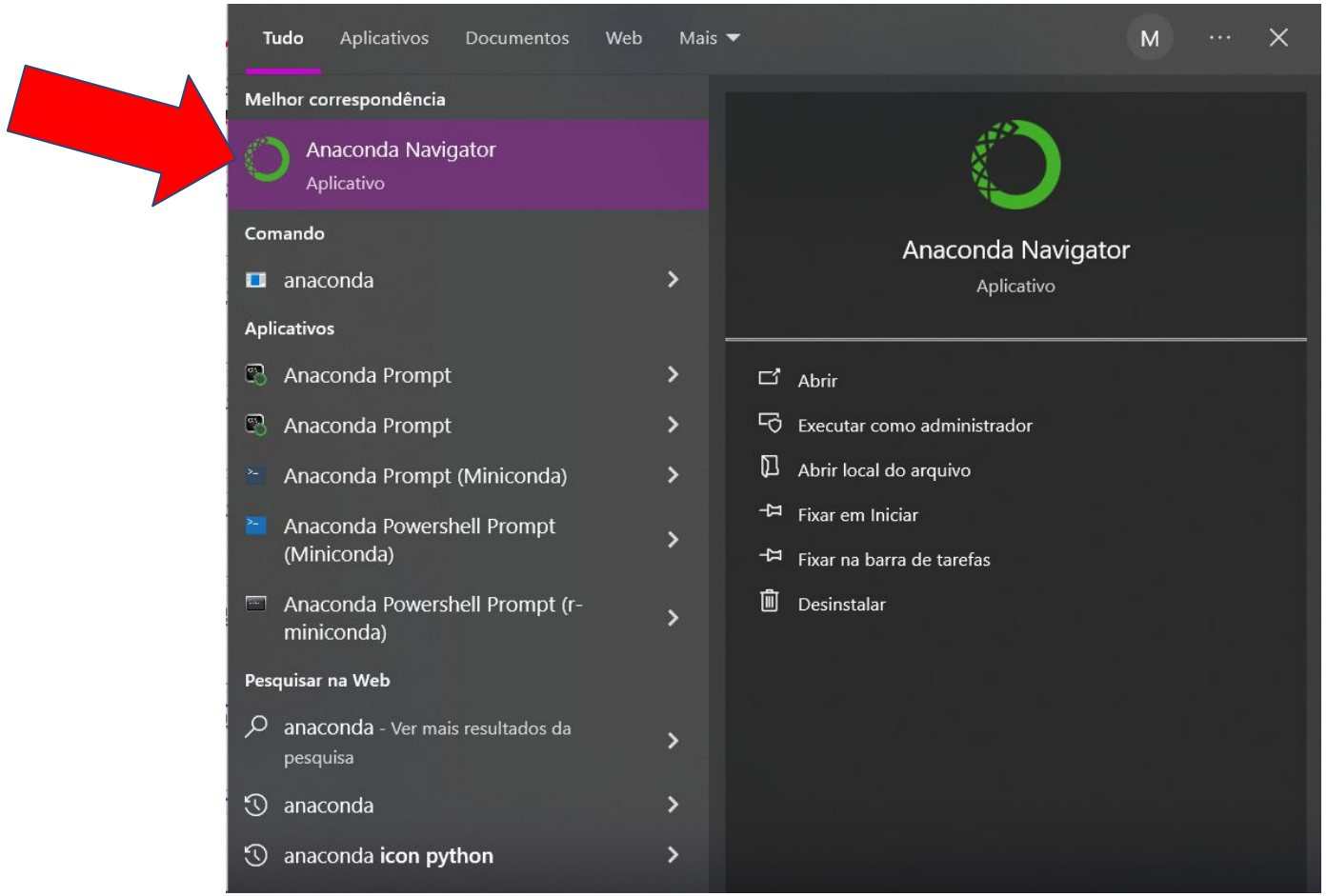
COMO BAIXAR?

Clique na opção **Register Anaconda3 as my default Python 3.13** e depois em **Install.** Após isso basta aguardar.



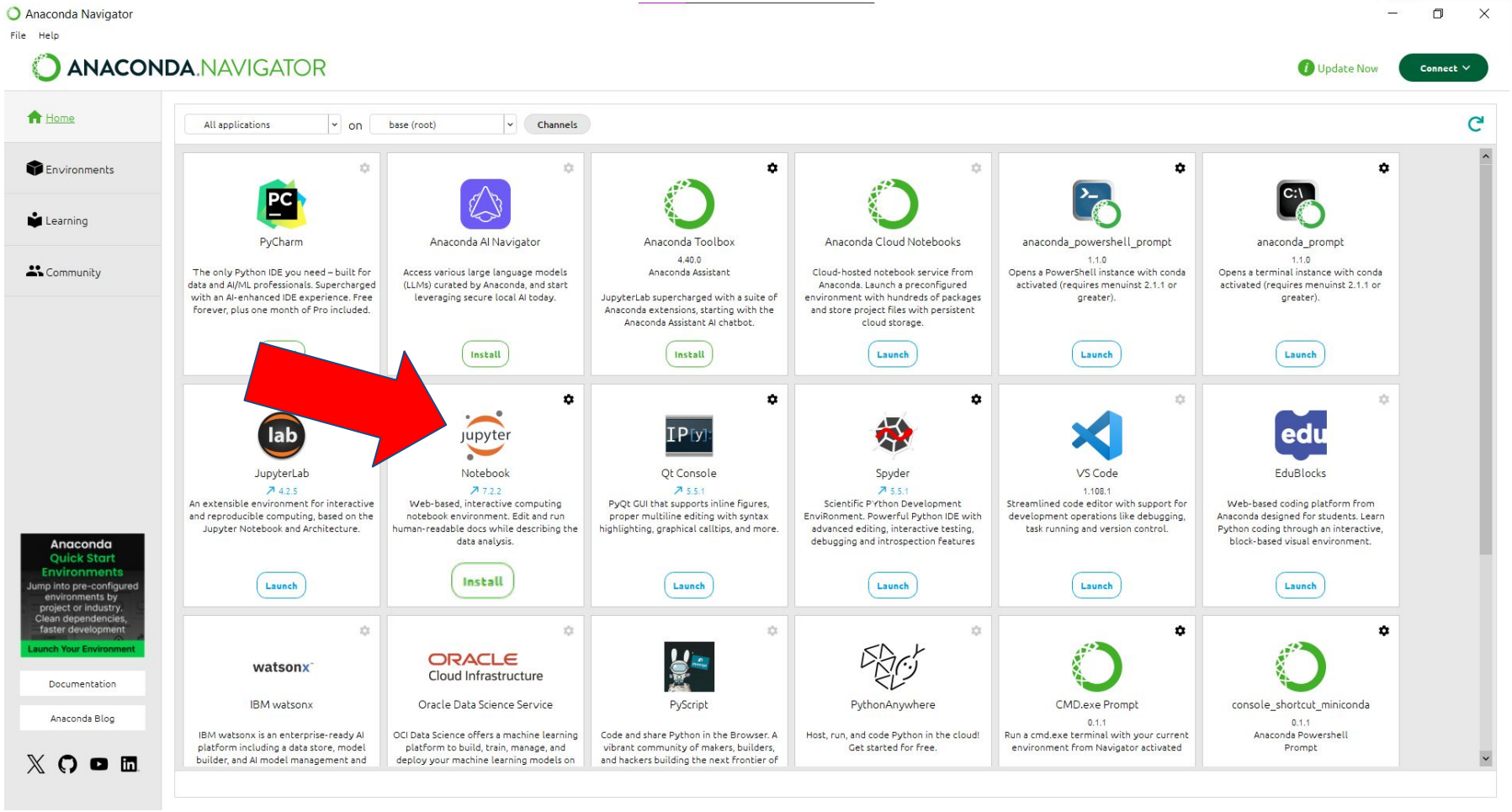
COMO ABRIR O ANACONDA?

Pesquise **Anaconda** na aba de pesquisa de aplicativos do seu sistema operacional e clique em **Anaconda Navigator**



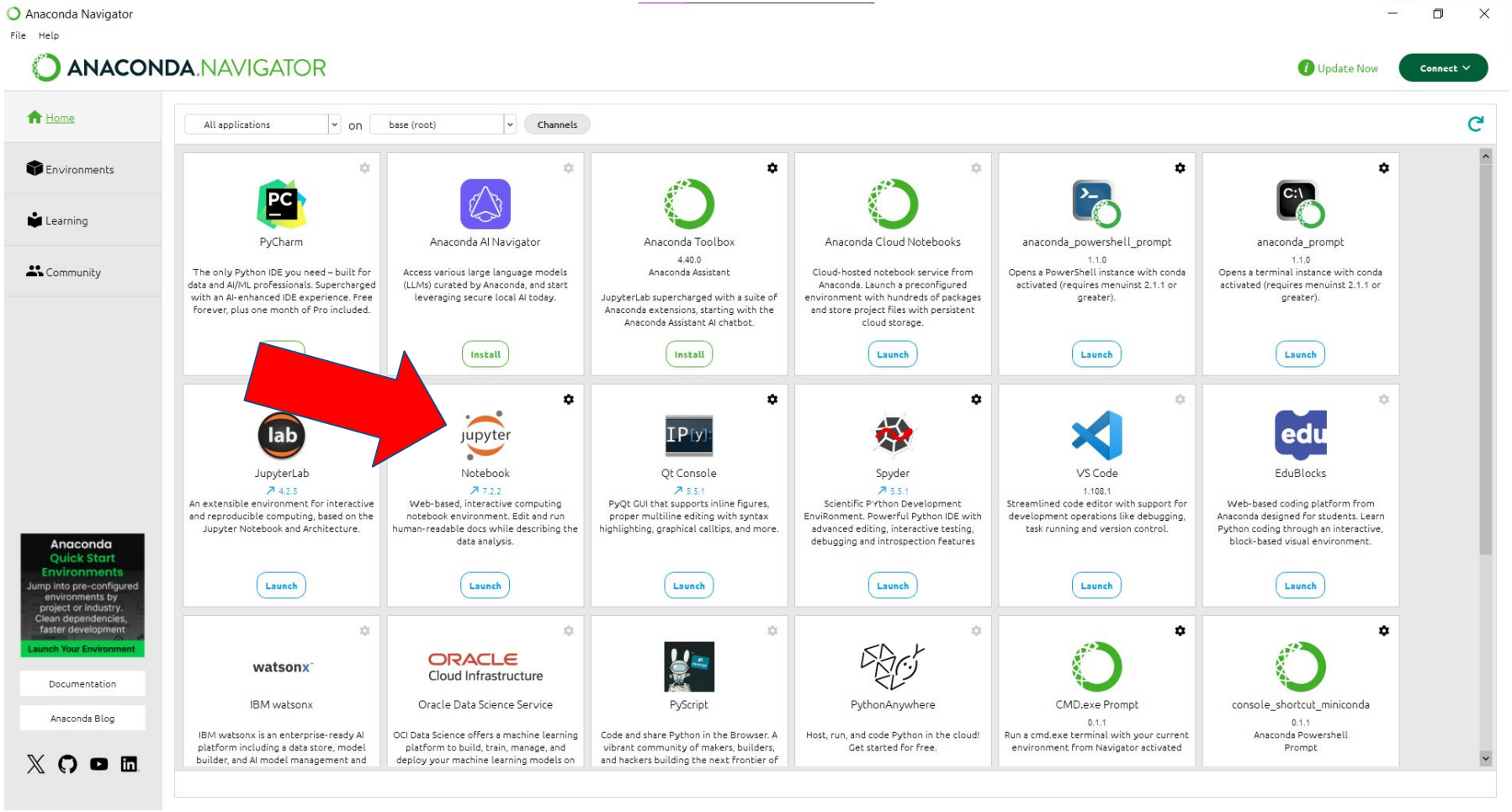
COMO ABRIR O JUPYTER NOTEBOOK?

É necessário instalar o jupyter notebook, clique em **Install**



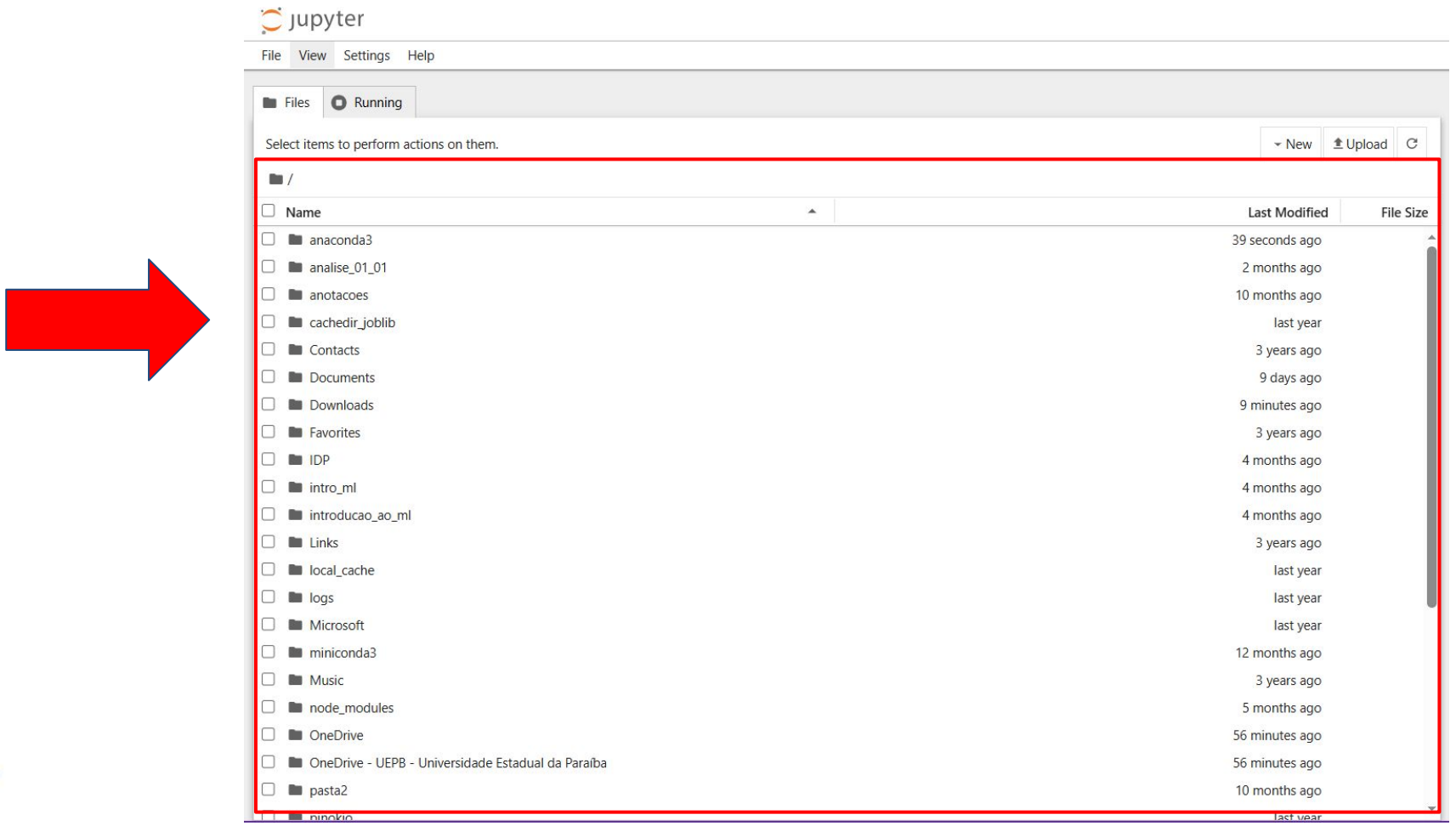
COMO ABRIR O JUPYTER NOTEBOOK?

Depois clique em **Launch,** ele abrirá no seu navegador padrão ou pedirá para você selecionar um navegador para abrir.



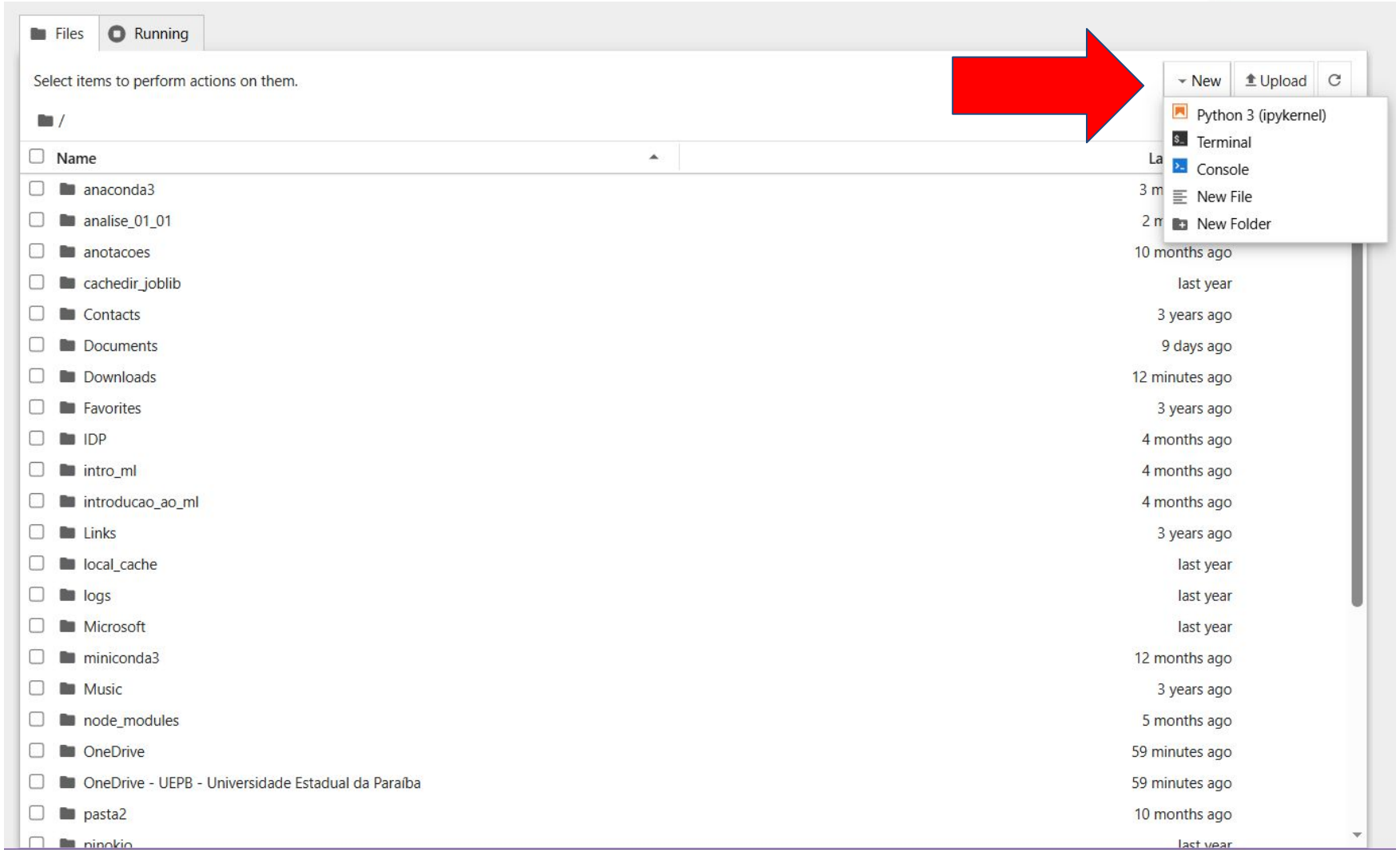
COMO USAR O JUPYTER NOTEBOOK?

Essa é o início do Jupyter Notebook. Aqui são os diretórios (pastas) do seu computador



COMO USAR O JUPYTER NOTEBOOK?

Clique em **New** e depois em **New Folder**

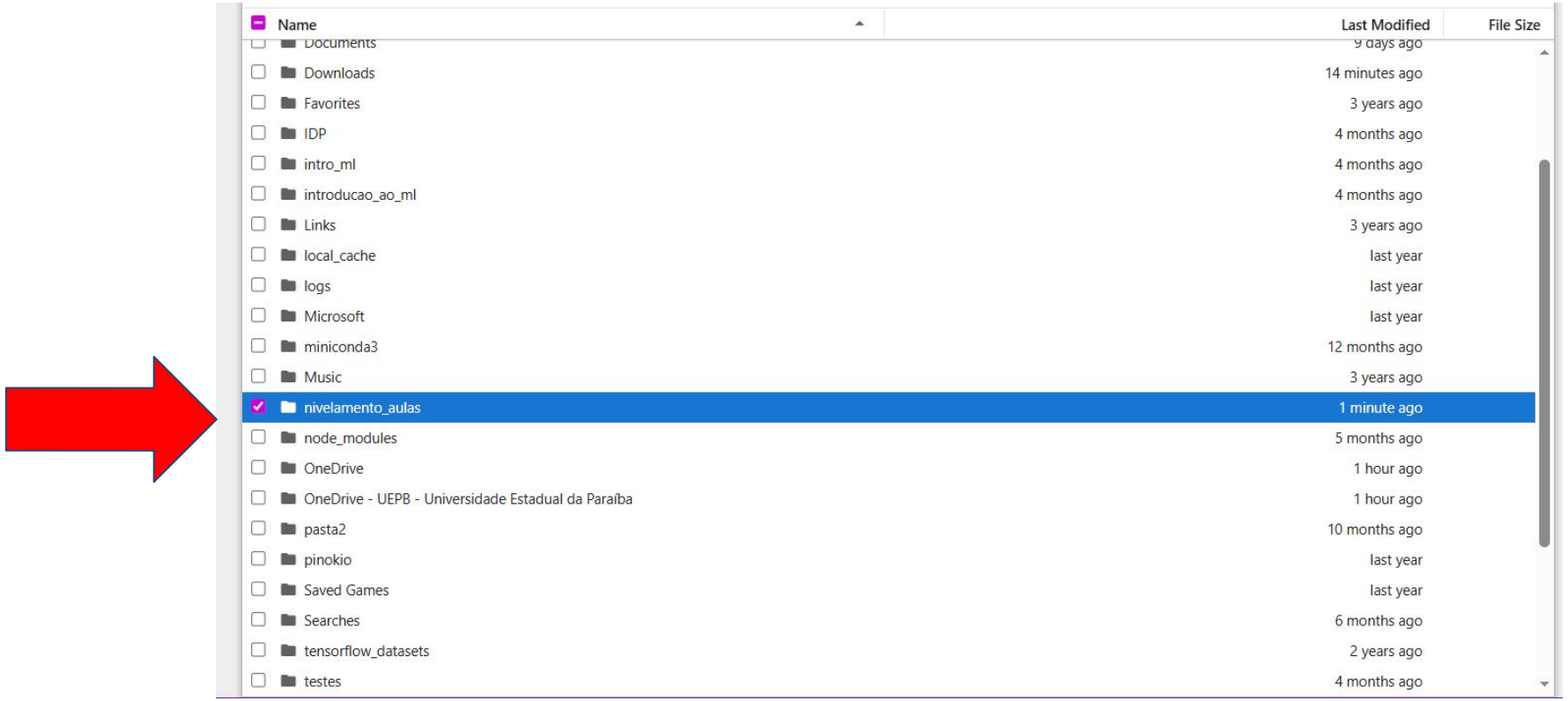


COMO USAR O JUPYTER NOTEBOOK?

No final dos diretórios, ele **criará um novo diretório para você.** Vamos renomear para **nivelamento_aulas**

COMO USAR O JUPYTER NOTEBOOK?

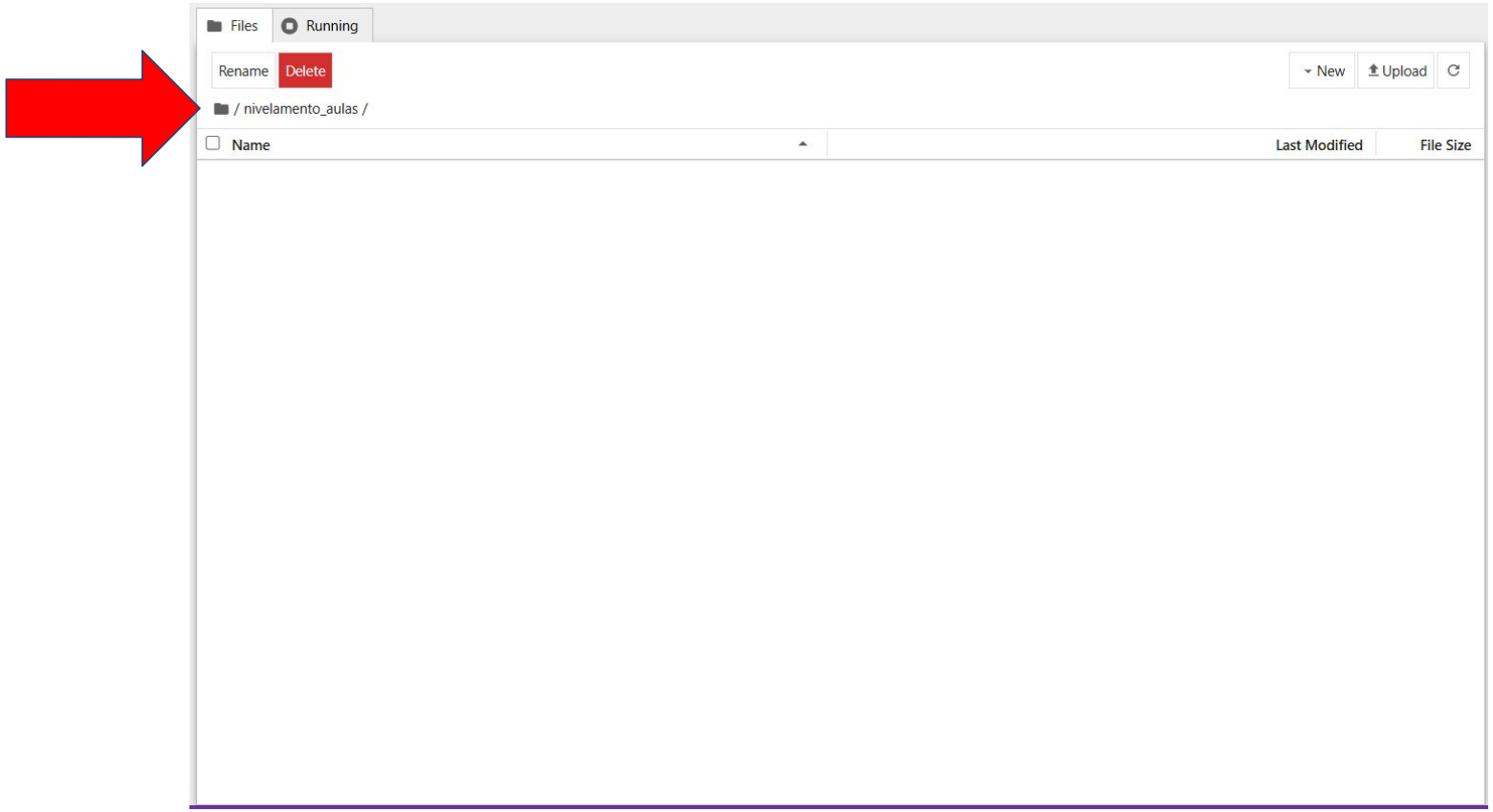
Dê dois cliques na nova pasta criada para entrar nela.



COMO USAR O JUPYTER NOTEBOOK?

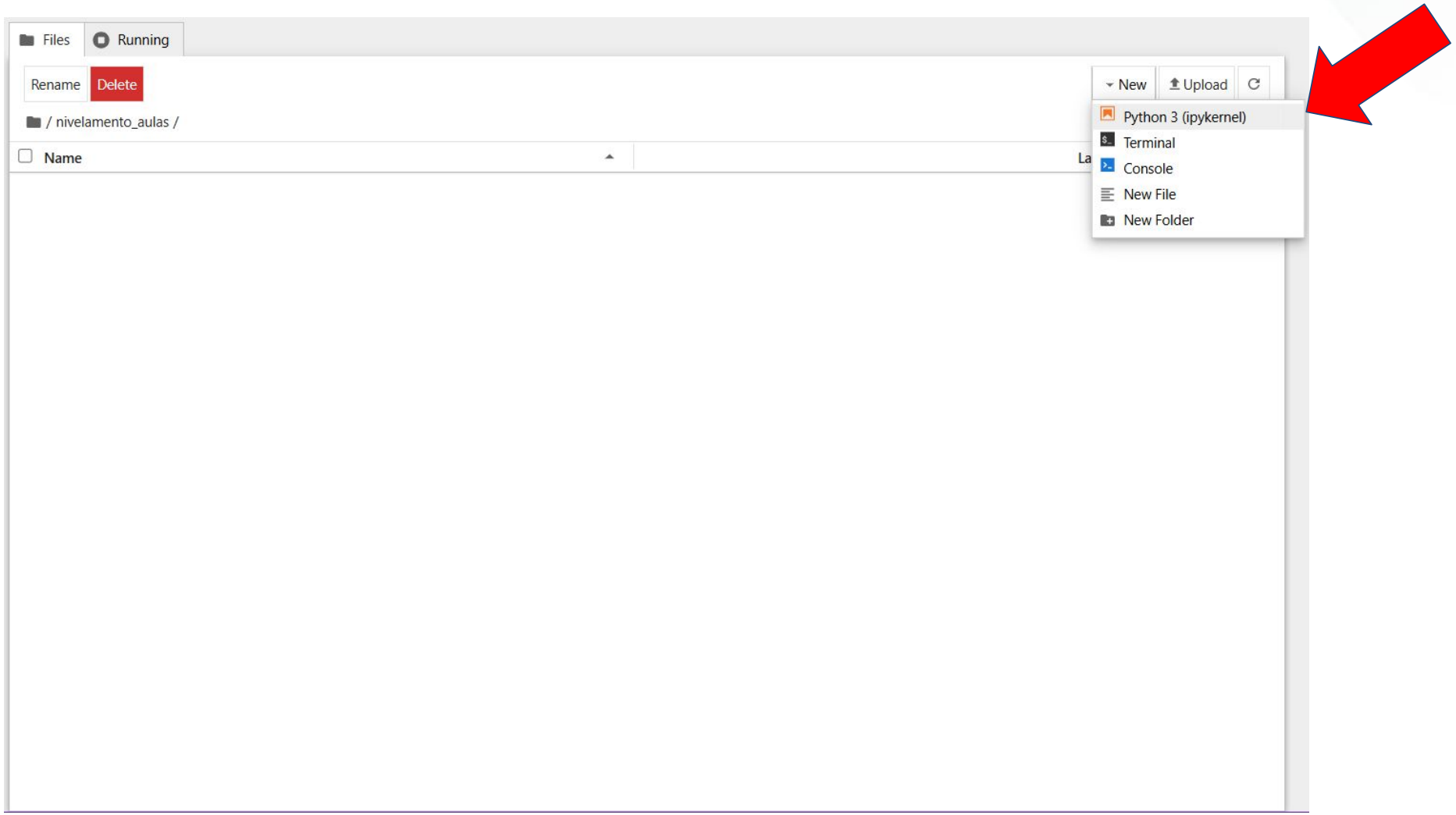
Ela estará vazia, note que no canto superior esquerdo dá pra ver que estamos dentro da pasta.

Caso queira voltar para a página inicial, clique no ícone da **pastinha** do lado do nome da pasta **nivelamento_aulas.**



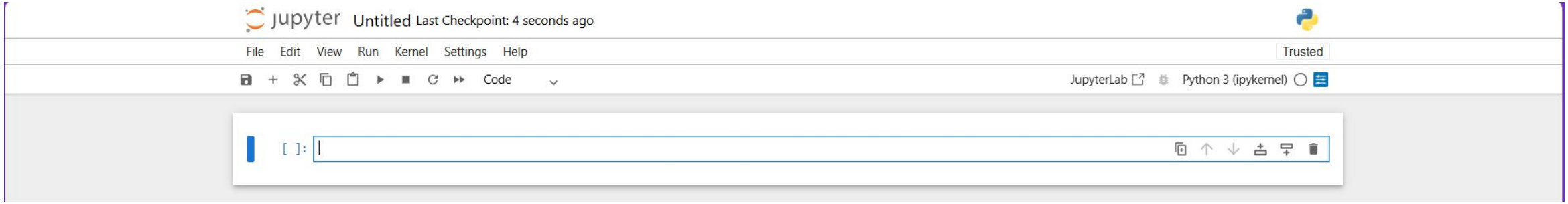
COMO CRIAR UM ARQUIVO JUPYTER?

Agora clique em New novamente e depois em **Python 3 (ipykernel)**



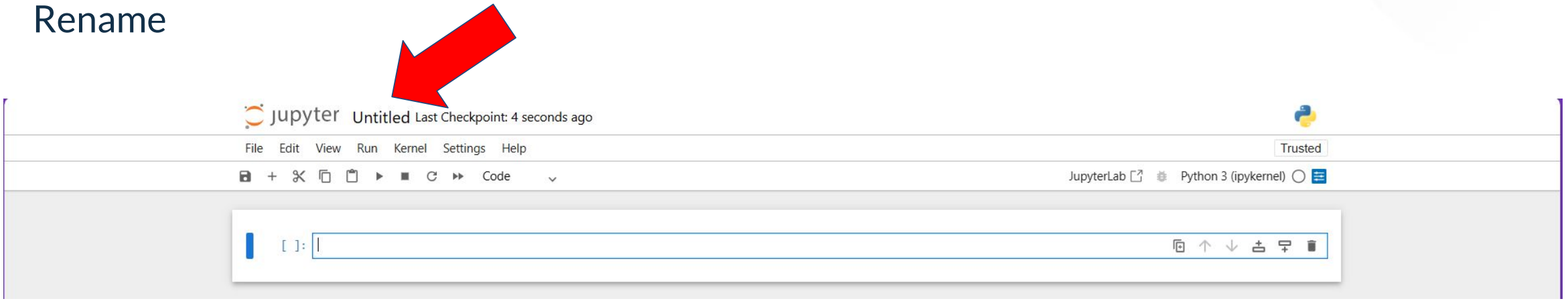
COMO CRIAR UM ARQUIVO JUPYTER?

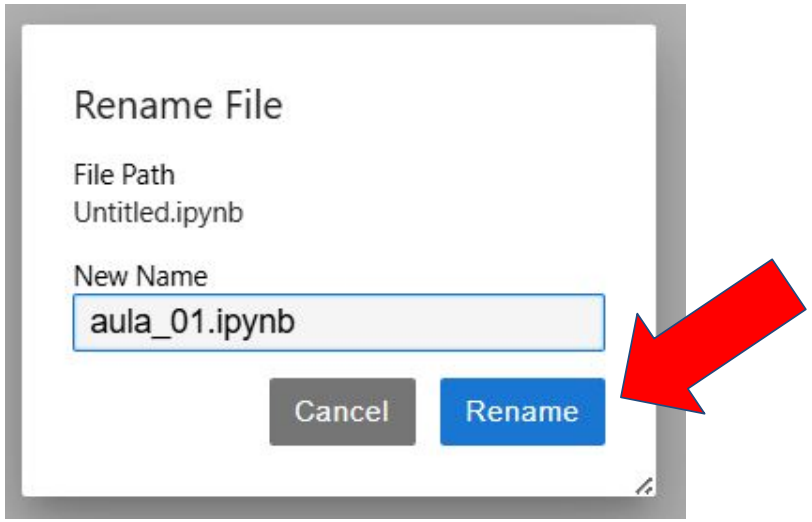
O jupyter notebook irá abrir outra aba contendo o modo editor.



COMO CRIAR UM ARQUIVO JUPYTER?

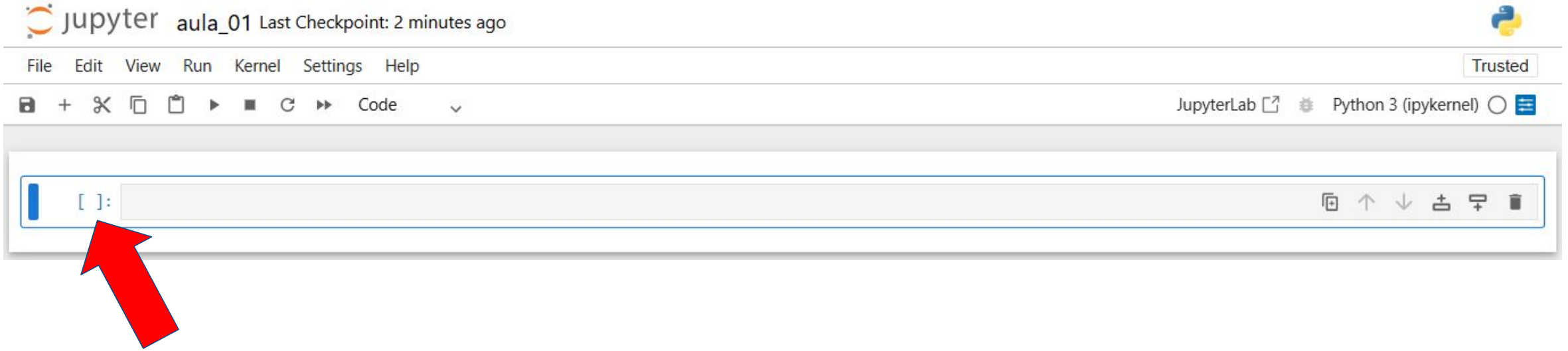
Vamos renomear o arquivo para **aula_01**, para isso clique no nome **Untitled.** Por fim clique em Rename





COMO CRIAR UM ARQUIVO JUPYTER?

O jupyter notebook funciona no esquema de células, uma célula pode ser para escrever código ou escrever texto. **Células com o [ ]:** são para escrever código Python.

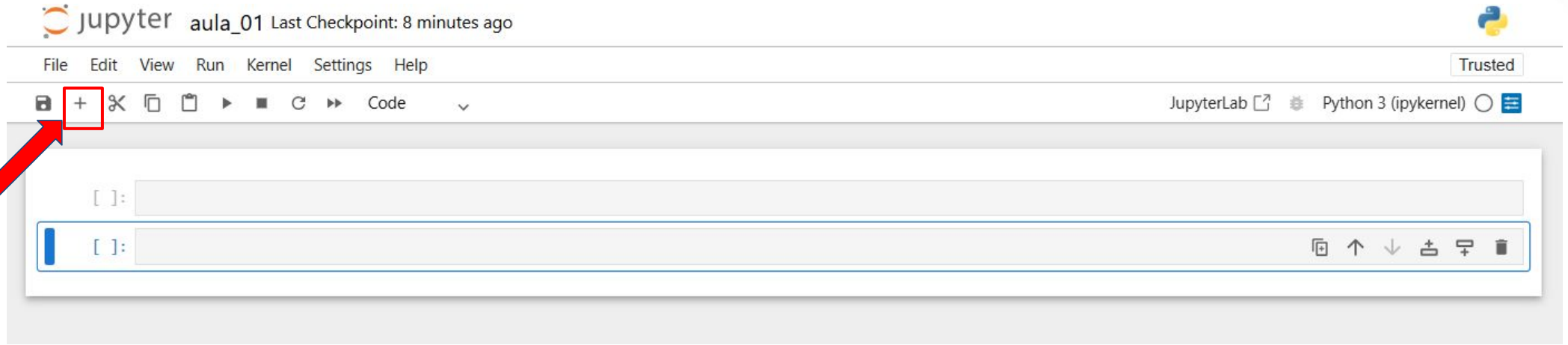


COMO CRIAR UM ARQUIVO JUPYTER?

Para modificar o tipo da célula, basta ir em **Code** e depois **Markdown**.

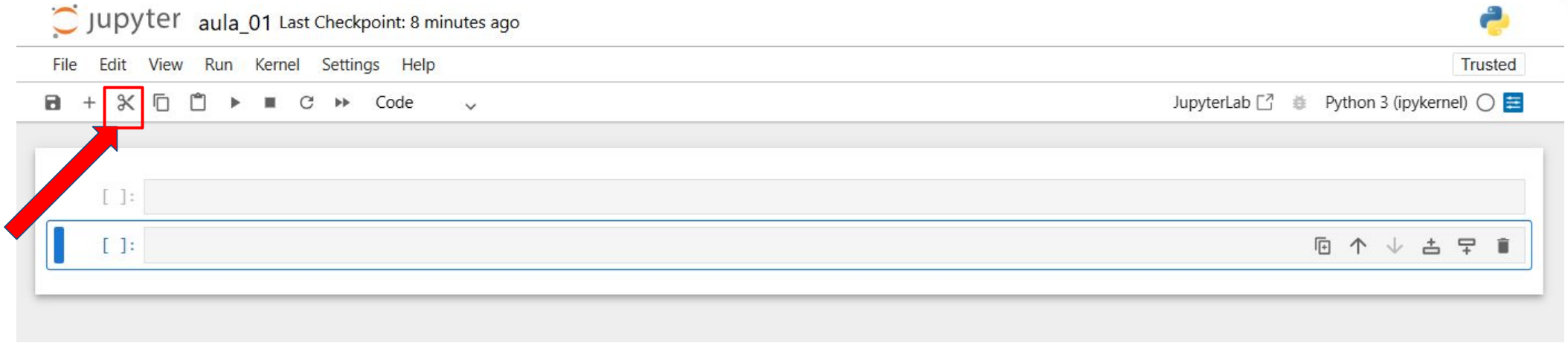
COMO CRIAR UM ARQUIVO JUPYTER?

Para inserir uma nova célula, basta clicar no ícone de **+.**



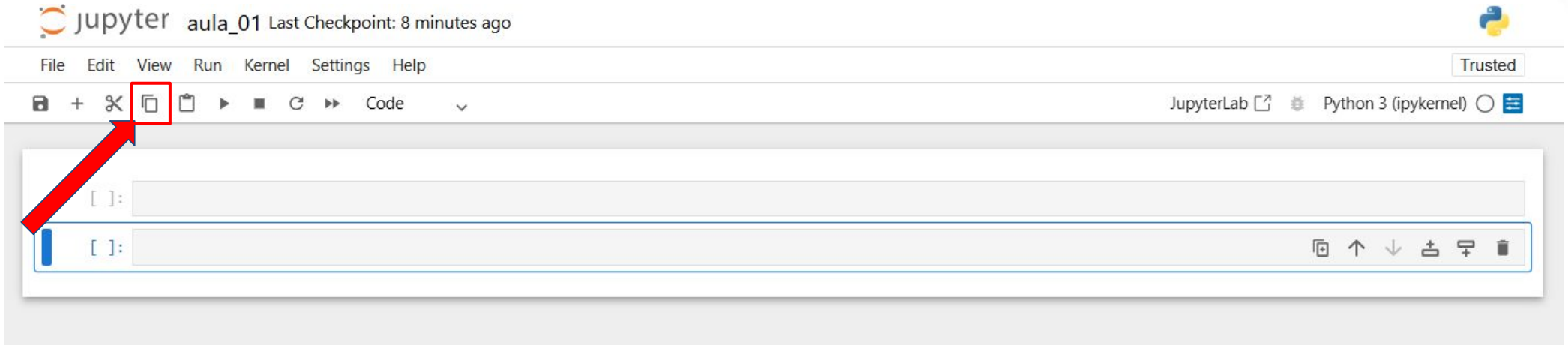
COMO CRIAR UM ARQUIVO JUPYTER?

Para remover uma célula, basta clicar no ícone da **tesoura.**



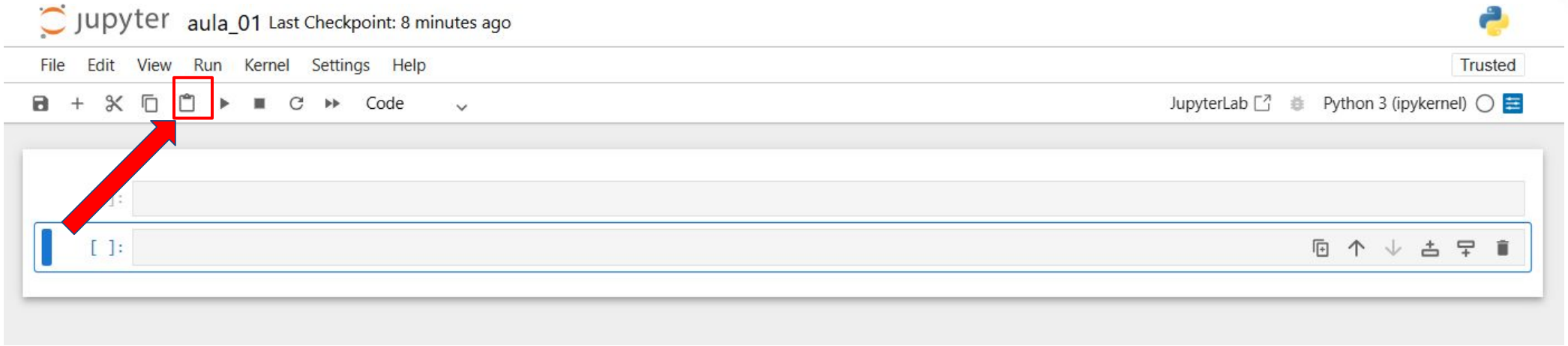
COMO CRIAR UM ARQUIVO JUPYTER?

Para copiar uma célula, basta clicar no ícone dos **papéis.**



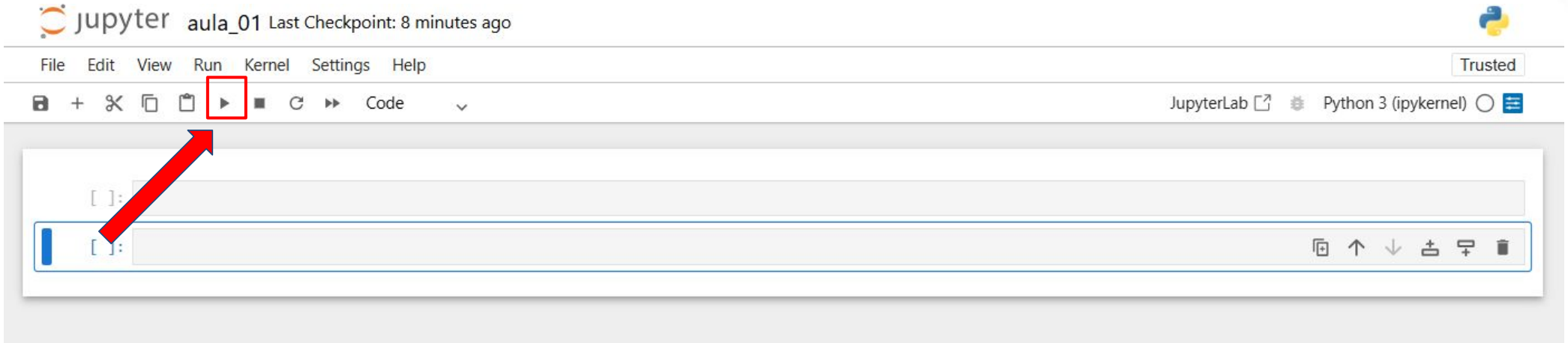
COMO CRIAR UM ARQUIVO JUPYTER?

Para colar uma célula, basta clicar no ícone da **prancheta.**



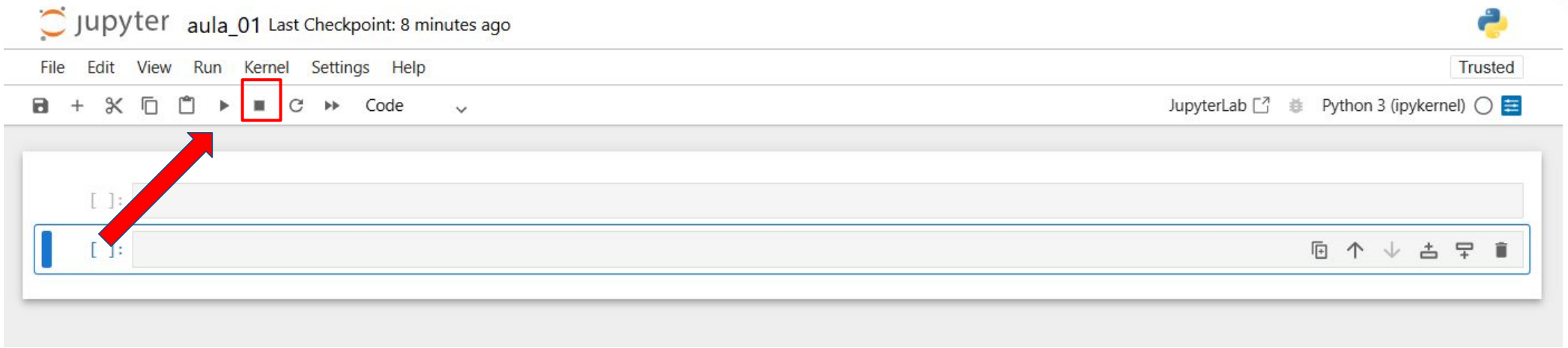
COMO CRIAR UM ARQUIVO JUPYTER?

Você pode executar os comandos dentro de uma célula com a **seta**



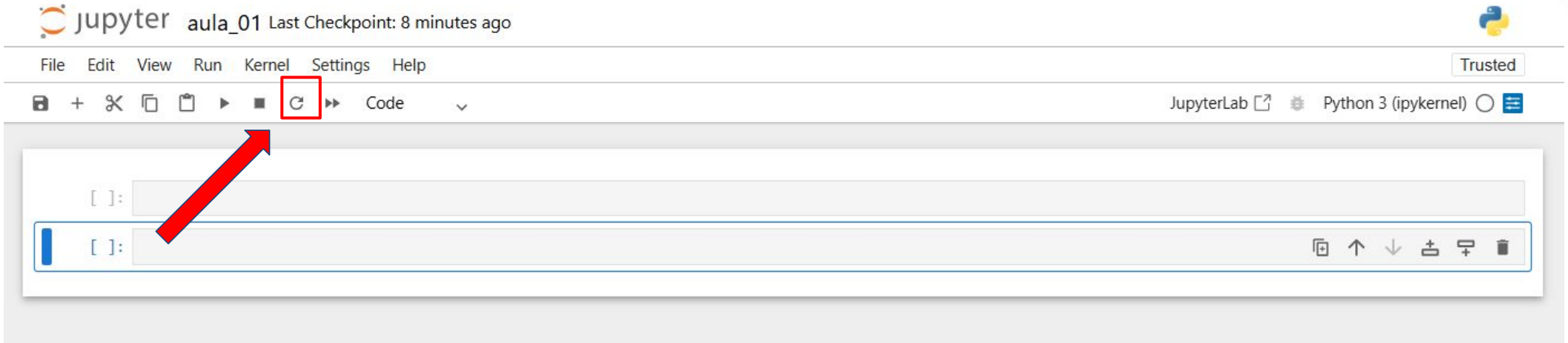
COMO CRIAR UM ARQUIVO JUPYTER?

Você pode parar a execução de alguma célula com o ícone do **quadradinho**



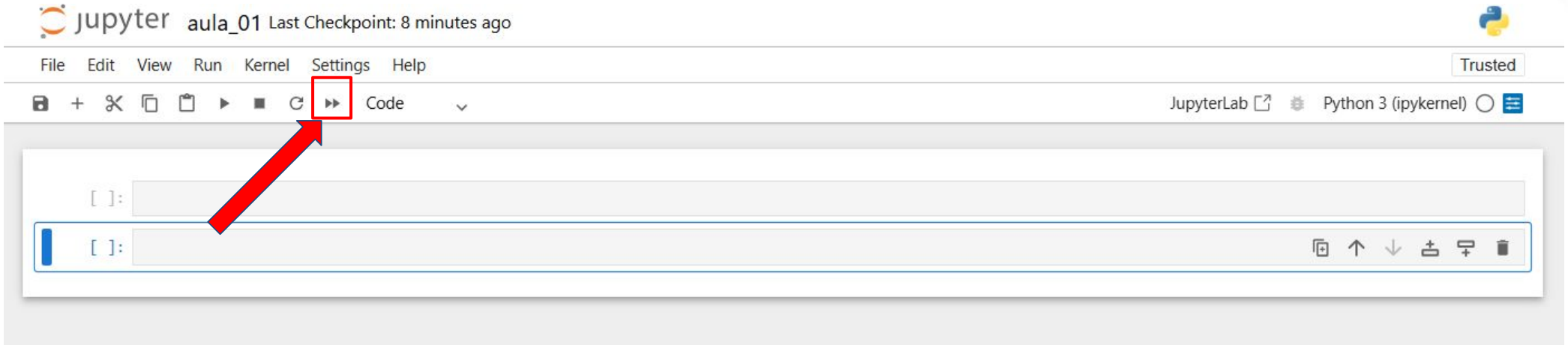
COMO CRIAR UM ARQUIVO JUPYTER?

Para reiniciar todo o ambiente jupyter, basta clicar na **seta circular**



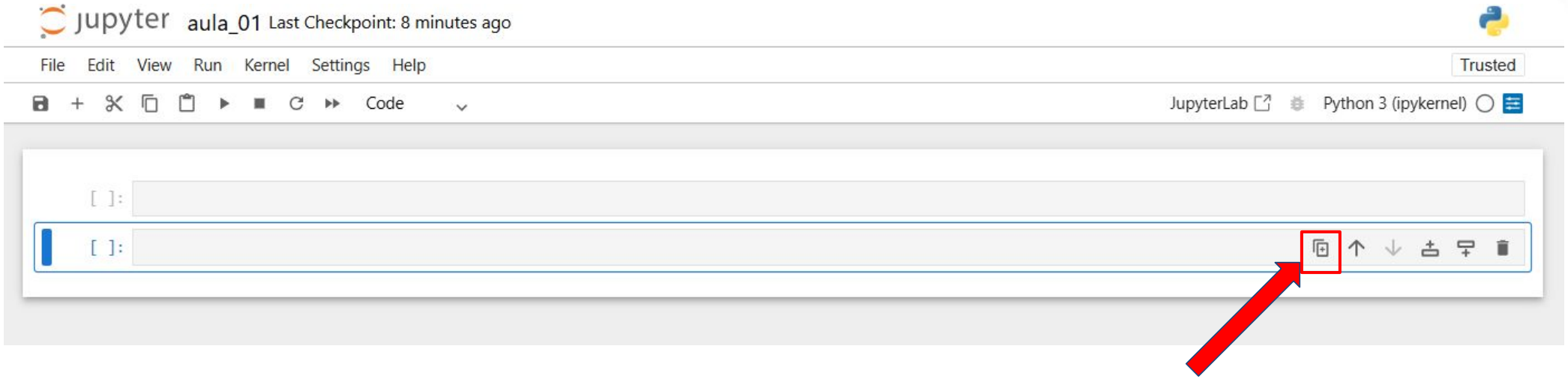
COMO CRIAR UM ARQUIVO JUPYTER?

Para executar TODAS as células, basta clicar no ícone das **duas setinhas**.



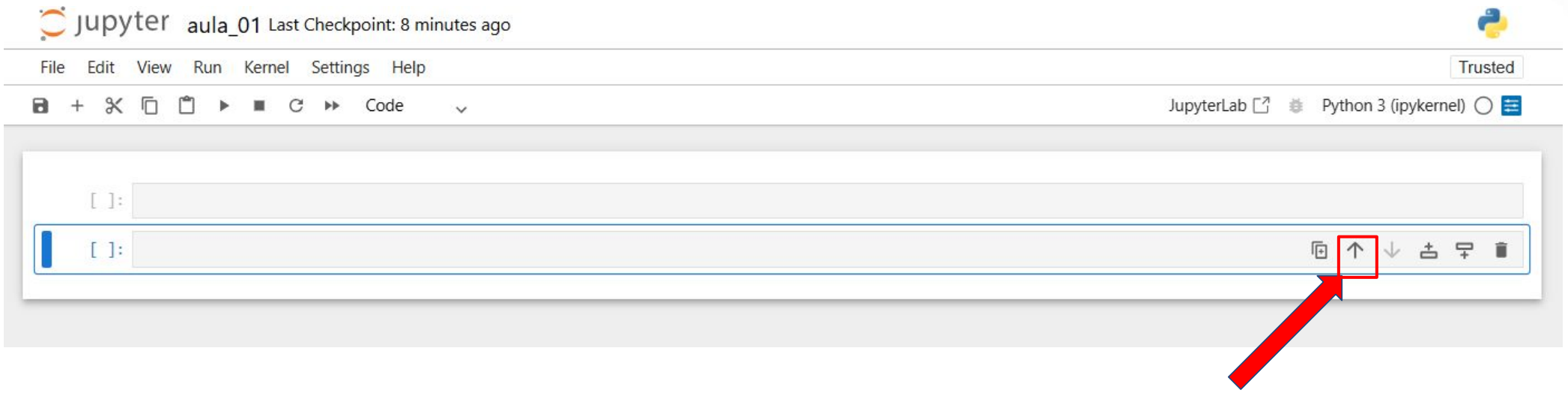
COMO CRIAR UM ARQUIVO JUPYTER?

Use o comando ao lado da célula para duplicá-la.



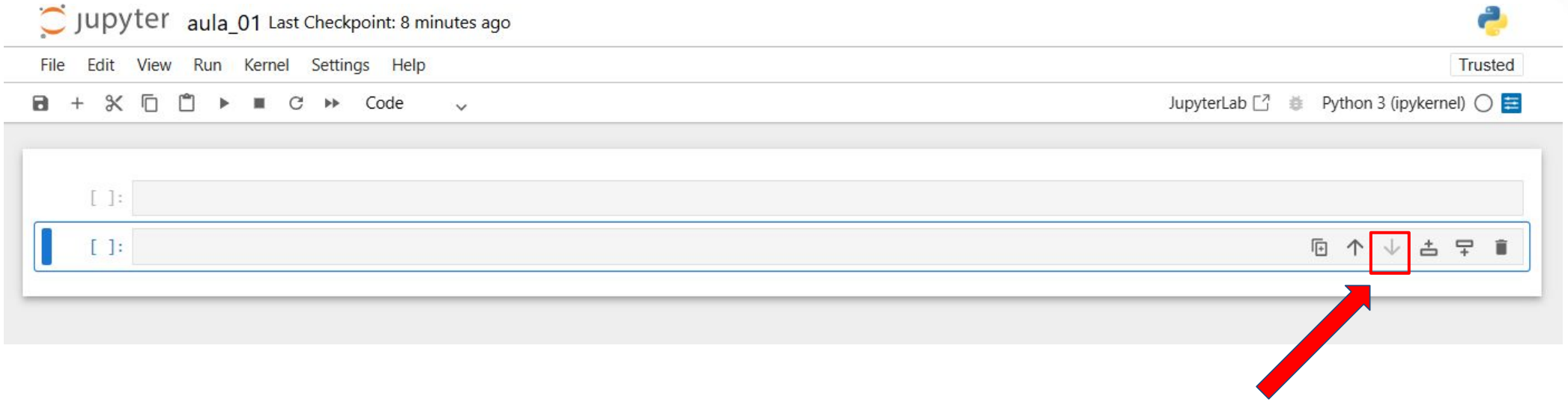
COMO CRIAR UM ARQUIVO JUPYTER?

A **seta pra cima** serve para mudar sua posição deixando ela acima.



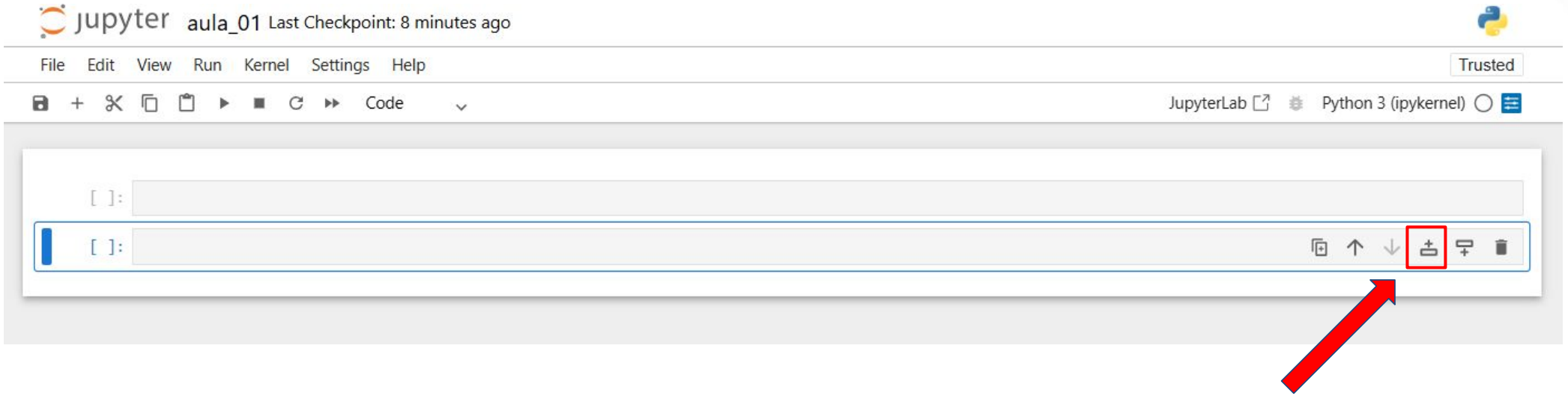
COMO CRIAR UM ARQUIVO JUPYTER?

A **seta pra baixo** serve para mudar sua posição deixando ela embaixo.



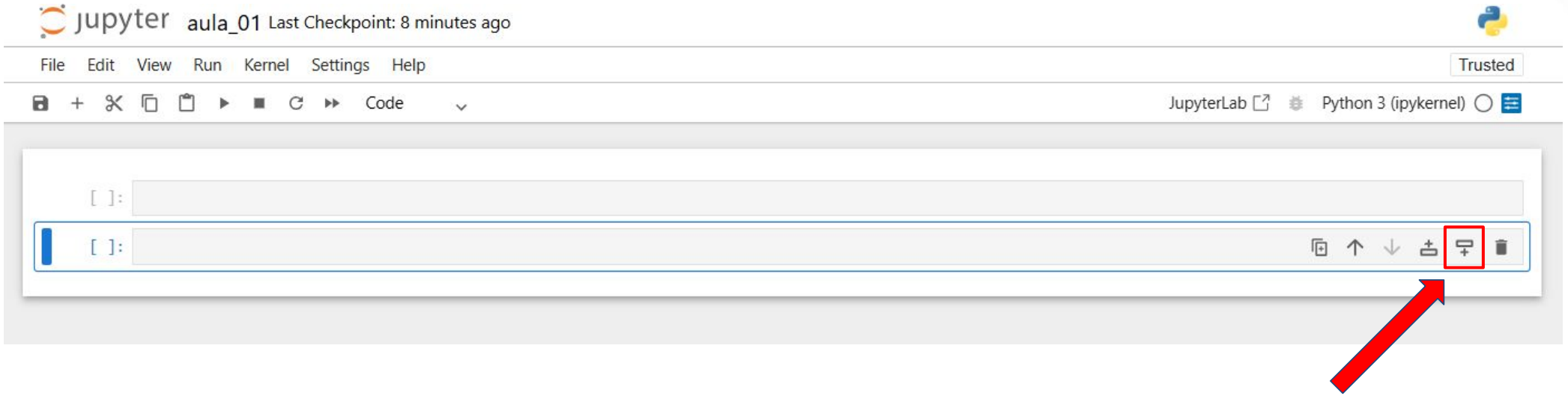
COMO CRIAR UM ARQUIVO JUPYTER?

Esse comando serve para **criar uma célula acima.**



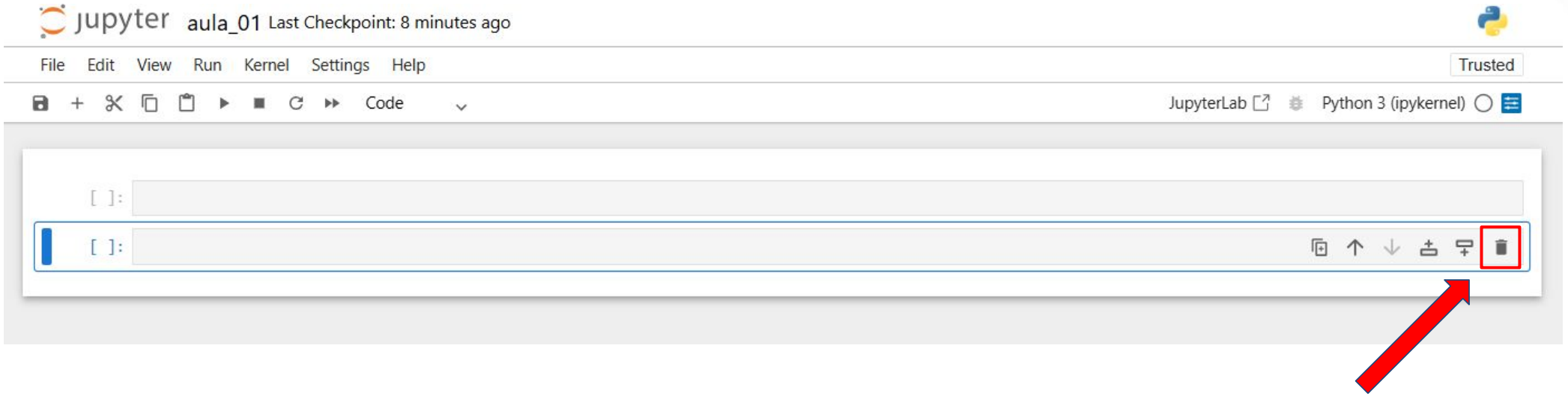
COMO CRIAR UM ARQUIVO JUPYTER?

Esse comando serve para **criar uma célula abaixo.**



COMO CRIAR UM ARQUIVO JUPYTER?

Esse comando serve para **remover a célula atual.**



Introdução às Bibliotecas para ciência de dados

Como manipular dados no python

RECAPITULANDO

- Na aula passada vimos a **diferença entre editores de texto, IDEs e o ambiente jupyter notebook com o Anaconda**, onde iremos utilizar a linguagem de programação python para realizar manipulações nos dados.
- Aprendemos como **baixar, instalar o Anaconda, abrir o Jupyter notebook e manipular o python.**

RECAPITULANDO

Vimos que o Python possui uma enorme quantidade de bibliotecas para manipular os dados.

Disponível em: [20+ Python Libraries for Data Science Professionals [2025 Edition] – Quantum™ Ai Labs]

O QUE É UMA BIBLIOTECA?

- Bibliotecas são coleções de módulos e funções pré-escritas que permitem aos desenvolvedores realizar tarefas complexas sem precisar escrever todo o código do zero.
- Isto é, basicamente são um conjunto de códigos escritos por pessoas que servem para resolver algum problema específico.
- O uso de bibliotecas promove a reutilização de código, o que acelera drasticamente o ciclo de desenvolvimento de qualquer projeto de software.
- Ao utilizar soluções que já foram testadas e validadas por milhares de outros programadores, você minimiza a ocorrência de bugs e foca seus esforços apenas na lógica de negócio exclusiva da sua aplicação.

O QUE É UMA BIBLIOTECA?

- Existem bibliotecas para quase todas as finalidades imagináveis, o que torna o Python uma linguagem extremamente versátil e poderosa no mercado atual.

- Temos o Pandas para análise de dados, o Django para desenvolvimento web robusto e o Matplotlib para criação de gráficos, permitindo que o desenvolvedor transite entre diferentes áreas com facilidade.

- A maioria das bibliotecas Python é de código aberto (Open Source), o que significa que são constantemente atualizadas e aprimoradas por uma comunidade global de desenvolvedores.

BIBLIOTECA PARA CIÊNCIA DE DADOS

Numpy

- O NumPy é a biblioteca fundamental para computação científica, servindo de base para quase todas as outras ferramentas de dados.
- É utilizado por debaixo dos panos na maioria das bibliotecas de ciência de dados como o pandas, matplotlib, seaborn, etc.
- Amplamente utilizado para operações aritméticas com matrizes e vetores.

Para mais informações, acesse: https://numpy.org/

BIBLIOTECA PARA CIÊNCIA DE DADOS

Pandas

- Considerada a "planilha de Excel" dentro do código, o Pandas fornece estruturas chamadas DataFrames, que permitem ler, filtrar, limpar e transformar dados tabulares com facilidade.
- É a ferramenta essencial para o processo de ETL (Extração, Transformação e Carga), permitindo lidar com dados ausentes, agrupar informações complexas e importar arquivos em formatos como CSV, JSON ou SQL

Para mais informações, acesse:https://pandas.pydata.org/

BIBLIOTECA PARA CIÊNCIA DE DADOS

Matplotlib e Seaborn

- Estas bibliotecas são responsáveis por transformar números em insights visuais através de gráficos.
- quanto o Matplotlib oferece controle total sobre cada detalhe do gráfico (como eixos e legendas), o Seaborn funciona como uma camada superior que facilita a criação de visualizações mais atraentes e complexas, como mapas de calor e gráficos de dispersão, com menos linhas de código.

Para mais informações, acesse:https://matplotlib.org/ Ou https://seaborn.pydata.org/

BIBLIOTECA PARA CIÊNCIA DE DADOS

Scikit-Learn

- O Scikit-Learn é a biblioteca padrão ouro para aprendizado de máquina em Python, oferecendo ferramentas simples e eficientes para análise preditiva.

- Ela inclui algoritmos para classificação, regressão e agrupamento, além de funcionalidades para pré-processamento de dados e avaliação de modelos, permitindo que desenvolvedores criem modelos de Machine Learning sem precisar implementar manualmente cálculos estatísticos pesados.

Para mais informações, acesse:https://scikit-learn.org/stable/index.html

Botando a Mão na Massa

Entendendo o problema de negócio

ANTES DE COMEÇAR...

- Entendemos que o Python possui um ecossistema de bibliotecas que são códigos que foram criados por outras pessoas para lidarem com ciência de dados.
- Mas antes de qualquer tipo de manipulação, é necessário saber qual o problema que você quer resolver.
- Já que não adianta extrair informações de uma base de dados sem contexto algum.

PROBLEMA INICIAL

Contextualizando

- Recebemos uma base bruta do Banco Mundial com indicadores de 2019. Para realizar uma análise de viabilidade de investimentos, não precisamos de todas as colunas, nem de todos os países.

- Precisamos reorganizar a base de dados para focar em nações que possuem perfis específicos e transformar dados técnicos em informações estratégicas para identificar líderes e lanternas globais em cada indicador.

PROBLEMA INICIAL

Perguntas a Serem Respondidas

- Como mostrar a base de dados dentro do ambiente de desenvolvimento?
- Quantos países e quantas variáveis temos no nosso banco de dados?
- Como calcular as estatísticas descritivas das variáveis?

PROBLEMA INICIAL

Perguntas a Serem Respondidas

- Como criar uma nova base de dados contendo apenas as colunas de riqueza e educação filtrando apenas países que possuem uma população superior a 100 milhões de habitantes?

- Como extrair o valor exato do PIB per capita da 10ª linha do DataFrame e, em seguida, selecionar as 5 primeiras linhas apenas para as colunas de terra arável e indústria?

PROBLEMA INICIAL

Perguntas a Serem Respondidas

- Como criar uma nova coluna que classifique como "Alta Escolaridade" países com Mão de Obra Qualificada acima de 50% e "Baixa Escolaridade" os demais?

- Como podemos localizar o país que possui a maior Carga Industrial e aquele com a menor área de terra arável do mundo?

- Quais países possuem uma porcentagem da força de trabalho com educação avançada superior a 25% e, simultaneamente, um gasto público em educação acima de 5% do PIB?

PROBLEMA INICIAL

Variáveis

- **NY.GDP.PCAP.CD (GDP_PC):** PIB per capita em dólares atuais. Representa a riqueza média produzida por cada habitante do país.
- **SL.TLF.ADVN.ZS (forca_trab_educ):** Porcentagem da força de trabalho que possui educação avançada (nível superior ou pós-graduação).
- **AG.LND.ARBL.HA.PC (arable_land):** Hectares de terra arável (cultivável) disponíveis por pessoa na população.
- **NV.IND.TOTL.ZS (industria_PERCPIB):** Porcentagem do PIB que vem da indústria (incluindo mineração, manufatura, construção e serviços de utilidade pública).

PROBLEMA INICIAL

Variáveis

- **SE.XPD.TOTL.GD.ZS (gasto_educ_PERCPIB):** Total de gastos públicos em educação expressos como uma porcentagem do PIB do país.
- **NE.GDI.FTOT.CD (FBKF):** Formação Bruta de Capital Fixo. Mede o valor das aquisições de ativos fixos (máquinas, equipamentos e infraestrutura) em dólares.
- **SP.POP.TOTL (populacao):** População total de residentes, independente do status legal, no meio do ano de 2019.

JÁ PODEMOS BOTAR A MÃO NA MASSA E COMEÇAR A ANALISAR OS DADOS? A RESPOSTA É NÃO!

PRECISAMOS CRIAR UM DICIONÁRIO DE DADOS!

PROBLEMA INICIAL

Dicionário de Dados

- **O que é um Dicionário de Dados?** É um documento centralizado que contém os metadados sobre o seu conjunto de dados.
- Imagine-o como um manual de instruções que explica detalhadamente o que cada coluna representa, garantindo que qualquer pessoa entenda exatamente o que está sendo processado

PROBLEMA INICIAL

Dicionário de Dados

- **Por que criar um dicionário antes de codar?** A criação do dicionário é o passo que separa um "curioso de dados" de um "Cientista de Dados".
- Ele serve para evitar conclusões erradas (ex: confundir dólar com real), padronizar o entendimento entre os membros da equipe e facilitar a escolha dos tipos de dados corretos no Pandas (inteiros, decimais ou textos).

PROBLEMA INICIAL

Dicionário de Dados

Como criar o dicionário na prática? Para construir um dicionário robusto, estruturamos uma tabela com 5 pilares fundamentais:

- **Nome:** O nome exato da coluna.
- **Descrição:** O significado funcional daquela informação.
- **Tipo de Dado:** De acordo com a classificação estatística (Qualitativa nominal, ordinal, quantitativa discreta ou contínua)
- **Unidade de Medida:** Crucial para saber se falamos de USD, %, ou anos.
- **Valores Válidos:** O intervalo esperado (ex: 0 a 100 para taxas) ou categorias (ex: Infecciosa/Não-Transmissível).

PROBLEMA INICIAL

Classificação Estatística de Dados

Dados Quantitativos São variáveis que expressam quantidades e podem ser medidas em uma escala numérica.

- **Discretas**, que representam contagens de números inteiros (ex: número de médicos, número de leitos)
- **Contínuas**, que podem assumir qualquer valor em um intervalo, geralmente com casas decimais

PROBLEMA INICIAL

Classificação Estatística de Dados

Dados Qualitativos (Categóricos) São variáveis que expressam atributos, qualidades ou categorias, não possuindo um valor numérico intrínseco.

- **Nominais**, onde não existe uma ordem entre as categorias (ex: Nome da Doença, Gênero, Tipo de Tratamento)
- **Ordinais,** onde existe uma hierarquia lógica (ex: Faixa Etária, Nível de Escolaridade)

PROBLEMA INICIAL

Considere o exemplo abaixo

O CEP é uma variável quantitativa ou qualitativa?

- Resposta: Qualitativa nominal, já que não faz sentido realizar operações aritméticas com o CEP.

A idade uma variável discreta ou contínua?

- Resposta: Depende! Posso considerar discreta ou contínua dependendo do meu contexto de negócio.

PROBLEMA INICIAL

E variáveis de data?

Na estatística, data é simplesmente data!

- Ela parece se comportar com uma variável quantitativa pois conseguirmos calcular um intervalo entre datas, mas parece ter uma ordem cronológica. Porém também há a característica de ser cíclica.

Vamos praticar!

PROBLEMA INICIAL

O Custo do Tratamento "Ao abrir a planilha de custos de uma cirurgia cardíaca, você encontra o valor de $ 15.450,75. **Como você classificaria a variável?**

- **Quantitativa Contínua.** Por que? O valor possui centavos (casas decimais). O dinheiro é uma grandeza que pode ser medida em qualquer fração dentro de um intervalo.

PROBLEMA INICIAL

Capacidade da UTI "O diretor do hospital informa que a unidade possui exatamente 42 leitos disponíveis no momento. **Como você classificaria a variável?**

- **Quantitativa Discreta.** Por que? Representa uma contagem de unidades inteiras. Não existe "meio leito" físico na gestão hospitalar.

PROBLEMA INICIAL

O Estágio da Doença "No prontuário, a gravidade da condição do paciente está marcada como 'Moderada' (em uma escala de Leve, Moderada e Grave). **Como você classificaria a variável?**

- **Qualitativa Ordinal.** Por que? É um atributo (texto), mas que possui uma hierarquia e ordem lógica clara de intensidade.

PROBLEMA INICIAL

Identificação Regional "Para mapear a origem das infecções, o analista usa o código 70000-000 (CEP) para agrupar os casos. **Como você classificaria a variável?**

- **Qualitativa Nominal.** Por que? Apesar de ser formado por algarismos, ele serve apenas como um rótulo de localização. Não faz sentido somar dois CEPs ou tirar a média deles.

VAMOS A OBRA!

Criando o diretório, executando o Jupyter e desenvolvendo código

RECAPITULANDO

- Na aula passada, vimos a importância da estrutura do DataFrame e das Series.
- Aprendemos que o Pandas não organiza os dados apenas em tabelas visuais, mas em estruturas indexadas onde cada coluna (Series) possui um tipo de dado específico que dita o que podemos fazer com ela.

RECAPITULANDO

- Entendemos que identificar se um dado é uma variável quantitativa ou qualitativa é o primeiro passo para não cometer erros básicos, como tentar somar colunas de texto ou agrupar dados por valores contínuos sem critério.

- Comprendemos que, enquanto o loc nos dá a liberdade de filtrar por nomes de países ou estados (como as UFs da RAIS), o iloc é a nossa ferramenta de precisão cirúrgica para fatiar o DataFrame ou extrair amostras específicas.

Problemática

- Você recebeu o arquivo **rais_mulheres_2008.xlsx**, que é uma amostra riquíssima do mercado formal.
- No entanto, ela não contém dados macroeconômicos. Para uma análise completa, você precisa saber também o PIB per capita dos estados.
- O problema: Os dados do PIB estão em um DataFrame pequeno de 27 linhas (df_pib_pc), enquanto a RAIS tem milhares de linhas. Se você tentar copiar e colar, o erro é certo.

O QUE DEVEMOS FAZER AGORA?

Dicionário de dados da RAIS

Nome da Coluna	Descrição Funcional	Tipo de Dado	Unidade de Medida	Valores Válidos
uf	Sigla da Unidade da Federação (Chave da RAIS).	Qualitativa Nominal	Texto (Sigla)	AC, AL, ..., TO
UF	Sigla da Unidade da Federação (Chave do PIB).	Qualitativa Nominal	Texto (Sigla)	AC, AL, ..., TO
salario	Remuneração nominal da trabalhadora.	Quantitativa Contínua	Real (R$)	Valor > 0
PIB_PC	Riqueza média por habitante do estado em 2008.	Quantitativa Contínua	Real (R$)	Valor > 0
escolaridade	Grau de instrução formal da trabalhadora.	Qualitativa Ordinal	Categoria	SUP. COMP, SUP. INCOMP, MEDIO COMPL, MEDIO INCOMP, FUND COMPL, 6. A 9. FUND, 5.A CO FUND, ATE 5.A INC

Dicionário de dados PIB

Nome da Coluna	Descrição Funcional	Tipo de Dado	Unidade de Medida	Valores Válidos
UF	Sigla da Unidade da Federação.	Qualitativa Nominal	Texto (Sigla)	SP, MG, RJ, DF, etc.
PIB_PC	Produto Interno Bruto per capita da UF em 2008.	Quantitativa Contínua	Real (R$)	5.372 a 45.977

JOINS

- Um join é simplesmente uma forma de juntar duas tabelas usando algo que elas têm em comum. Note que:

BASE RAIS

Nome	UF	Escolaridade	Salário
Ana	SP	SUP. COMP	4.500
Beatriz	CE	MEDIO COMPL	2.100
Carla	SP	MEDIO COMPL	2.800
Daniela	DF	SUP. COMP	6.200
Elisa	MG	MEDIO INCOMP	1.900
Fernanda	XX	FUND COMPL	1.500

BASE PIB

UF	PIB_PC
CE	7.111
DF	45.977
SP	24.456
RJ	22.500
BA	8.900

JOINS

- O que tem em comum em ambas as tabelas?

BASE RAIS

Nome	UF	Escolaridade	Salário
Ana	SP	SUP. COMP	4.500
Beatriz	CE	MEDIO COMPL	2.100
Carla	SP	MEDIO COMPL	2.800
Daniela	DF	SUP. COMP	6.200
Elisa	MG	MEDIO INCOMP	1.900
Fernanda	XX	FUND COMPL	1.500

BASE PIB

UF	PIB_PC
CE	7.111
DF	45.977
SP	24.456
RJ	22.500
BA	8.900

JOINS

- Note que o que desejo é que a **BASE RAIS** contenha a informação da **BASE PIB**, portanto preciso fazer o **left join**!

BASE RAIS

Nome	UF	Escolaridade	Salário
Ana	SP	SUP. COMP	4.500
Beatriz	CE	MEDIO COMPL	2.100
Carla	SP	MEDIO COMPL	2.800
Daniela	DF	SUP. COMP	6.200
Elisa	MG	MEDIO INCOMP	1.900
Fernanda	XX	FUND COMPL	1.500

LEFT JOIN

BASE PIB

UF	PIB_PC
CE	7.111
DF	45.977
SP	24.456
RJ	22.500
BA	8.900

=

Nome	UF	Escolaridade	Salário	PIB_PC
Ana	SP	SUP. COMP	4.500	24.456
Beatriz	CE	MEDIO COMPL	2.100	7.111
Carla	SP	MEDIO COMPL	2.800	24.456
Daniela	DF	SUP. COMP	6.200	45.977
Elisa	MG	MEDIO INCOMP	1.900	NULL
Fernanda	XX	FUND COMPL	1.500	NULL

JOINS

- E se quiséssemos apenas o que tem em comum em cada base de dados? Usaríamos o **inner join?**

BASE RAIS

Nome	UF	Escolaridade	Salário
Ana	SP	SUP. COMP	4.500
Beatriz	CE	MEDIO COMPL	2.100
Carla	SP	MEDIO COMPL	2.800
Daniela	DF	SUP. COMP	6.200
Elisa	MG	MEDIO INCOMP	1.900
Fernanda	XX	FUND COMPL	1.500

INNER JOIN

BASE PIB

UF	PIB_PC
CE	7.111
DF	45.977
SP	24.456
RJ	22.500
BA	8.900

=

Nome	UF	Escolaridade	Salário	PIB_PC
Ana	SP	SUP. COMP	4.500	24.456
Carla	SP	MEDIO COMPL	2.800	24.456
Beatriz	CE	MEDIO COMPL	2.100	7.111
Daniela	DF	SUP. COMP	6.200	45.977

JOINS

- E se quiséssemos que a BASE PIB contenha as informações da BASE RAIS? Nesse caso usaremos o **right join**!

BASE RAIS

Nome	UF	Escolaridade	Salário
Ana	SP	SUP. COMP	4.500
Beatriz	CE	MEDIO COMPL	2.100
Carla	SP	MEDIO COMPL	2.800
Daniela	DF	SUP. COMP	6.200
Elisa	MG	MEDIO INCOMP	1.900
Fernanda	XX	FUND COMPL	1.500

RIGHT JOIN

BASE PIB

UF	PIB_PC
CE	7.111
DF	45.977
SP	24.456
RJ	22.500
BA	8.900

=

Nome	UF	Escolaridade	Salário	PIB_PC
Ana	SP	SUP. COMP	4.500	24.456
Carla	SP	MEDIO COMPL	2.800	24.456
Beatriz	CE	MEDIO COMPL	2.100	7.111
Daniela	DF	SUP. COMP	6.200	45.977
NULL	RJ	NULL	NULL	22.500
NULL	BA	NULL	NULL	8.900

JOINS

- E se quiséssemos juntar tudo de informação? Nesse caso usaremos o **outer join**!

BASE RAIS

Nome	UF	Escolaridade	Salário
Ana	SP	SUP. COMP	4.500
Beatriz	CE	MEDIO COMPL	2.100
Carla	SP	MEDIO COMPL	2.800
Daniela	DF	SUP. COMP	6.200
Elisa	MG	MEDIO INCOMP	1.900
Fernanda	XX	FUND COMPL	1.500

OUTER JOIN

BASE PIB

UF	PIB_PC
CE	7.111
DF	45.977
SP	24.456
RJ	22.500
BA	8.900

=

Nome	UF	Escolaridade	Salário	PIB_PC
Ana	SP	SUP. COMP	4.500	24.456
Carla	SP	MEDIO COMPL	2.800	24.456
Beatriz	CE	MEDIO COMPL	2.100	7.111
Daniela	DF	SUP. COMP	6.200	45.977
Elisa	MG	MEDIO INCOMP	1.900	NULL
Fernanda	XX	FUND COMPL	1.500	NULL
NULL	RJ	NULL	NULL	22.500
NULL	BA	NULL	NULL	8.900

Problemática

Agora podemos analisar os dados para responder as seguintes perguntas:

- Como podemos descobrir quantas trabalhadoras existem por Unidade Federativa (UF) na nossa amostra e qual a porcentagem que cada estado representa em relação ao total do país?
- Crie um DataFrame chamado df_amostra contendo apenas 5 mulheres selecionadas aleatoriamente.

Problemática

- Crie um DataFrame chamado df_amostra contendo apenas 5 mulheres selecionadas aleatoriamente.
    - Se tentarmos unir essa amostra de 5 pessoas com a base de PIB (que tem 27 estados), qual comportamento você espera observar ao testar:
        - **Left Join:** Veremos as 5 mulheres com seus PIBs ou os 27 estados com PIBs vazios?
        - **Right Join:** O que acontece com os estados que não possuem nenhuma representante nessas 5 mulheres selecionadas?

VAMOS A OBRA!

Criando o diretório, executando o Jupyter e desenvolvendo código

O Problema do Banco Mundial

Até agora, conseguimos identificar o PIB de um país específico ou listar os 5 mais ricos. Mas, para um gestor global, restam perguntas que os dados brutos, sozinhos, não respondem de forma clara.

- **E se quiséssemos saber qual a população total em cada um dos quintis de riqueza?**
- **E se quiséssemos saber qual é o investimento médio em educação comparando por quintis?**
- **Qual o investimento mínimo, máximo, a mediana e a média comparando por quintis?**

A Resposta: Agregação de Dados

- Para responder a essas perguntas, não olhamos mais para o "País A" ou "País B", mas sim para o Grupo. É aqui que entra a Agregação:

- Agregar é o ato de "resumir" a informação. Em vez de termos 200 linhas de países, passamos a ter apenas 5 linhas (uma para cada quintil: pobres, médio pobres, médio, médio ricos e ricos).

A Resposta: Agregação de Dados

Por que isso é valioso?

- **Expressividade:** Com apenas uma linha de código, transformamos microdados dispersos em inteligência estratégica.
- **Visão Macro:** Conseguimos enxergar padrões que estão ocultos quando olhamos os dados linha por linha.

A Resposta: Agregação de Dados

Suponhamos que queremos saber quantas vendas cada cidade fez:

cidade	produto	vendas	quantidade
Recife	A	100	10
Recife	B	150	15
Salvador	A	200	20
Salvador	B	120	12
Fortaleza	A	180	18
Fortaleza	B	160	16

A Resposta: Agregação de Dados

Podemos agrupar cada venda da cidade com base na soma:

cidade	produto	vendas	quantidade
Recife	A	100	10
Recife	B	150	15
Salvador	A	200	20
Salvador	B	120	12
Fortaleza	A	180	18
Fortaleza	B	160	16

cidade	total_vendas
Fortaleza	340
Recife	250
Salvador	320

VAMOS A OBRA!

Criando o diretório, executando o Jupyter e desenvolvendo código

Agora é com vocês!

Usando a base de dados **rais_mulheres_2008.xlsx** . Resolva abaixo:

- Exiba as estatísticas descritivas (média, desvio padrão, quartis) da coluna de salário.
- Filtre o DataFrame para exibir apenas as trabalhadoras com Ensino Superior Completo (SUP. COMP).
- Conte quantas trabalhadoras existem em cada UF.
- Crie uma nova coluna chamada salario_anual que seja o valor do salário multiplicado por 12.
- Realize um Inner Join entre df_rais e df_pib_pc para anexar o PIB de cada estado às trabalhadoras. Nomeie essa base de base_faixa.

Agora é com vocês!

- Na base_faixa, crie uma coluna chamada faixa_pib que divida os valores de PIB_PC em 3 categorias (Baixo, Médio, Alto) usando quantis.
- Crie uma amostra aleatória de 10 mulheres e realize um Right Join com a tabela de PIB. Observe o que acontece com os estados que não possuem representantes na amostra.
- Agrupe os dados da base_faixa pela coluna escolaridade e calcule a média e a mediana dos salários.
- Agrupe os dados da base_faixa pela nova coluna faixa_pib e calcule a soma total da remuneração e a contagem de trabalhadoras por grupo.
- Faça um agrupamento múltiplo: calcule a média salarial cruzando faixa_pib e escolaridade simultaneamente na base_faixa.

Dúvidas?

Obrigado pela atenção!

Mateus Rocha

Cientista de Dados | Estatístico