Aula 15 - Mineração de dados Parte 1: Notebooks

Como é sabido a essa altura do campeonato, os dados são atualmente o artefato mais valioso do mundo, sendo considerado o "novo ouro", superando valores do próprio ouro e também do petróleo. Um dos motivos principais do Python ser a linguagem número 1 do mundo atualmente é a sua facilidade em minerar dados de diversas fontes, como sites, bancos de dados e arquivos **.csv**. Esse é um tipo de tarefa muito requisitado por instituições financeiras, como bancos, por exemplo, e é um dos pré-requisitos para a criação de sistemas de IA (inteligência artificial).

A partir daqui, iremos utilizar dois recursos muito utilizados para esse fim:

- Jupyter Notebooks
- Pandas

Notebooks - Jupyter Notebook, Jupyter Lab e Google Colaboratory

---

O que são os notebooks?

Os ***notebooks*** em análise de dados não são computadores do tipo *laptop*, mas sim verdadeiros "cadernos de anotações digitais", onde é possível inserir textos normais, *hypertextos* (HTML) e códigos-fonte. Por padrão os notebooks aceitam códigos de 3 linguagens de programação: Julia, R e Python. É claro que em nossas aulas, utilizaremos o Python, mas há a possibilidade de usar códigos de outras linguagens de programação nos notebooks também, como o JavaScript ou o C++, por exemplo, graças a algumas extensões. A grande vantagem de se trabalhar com os notebooks é a possibilidade não só de digitar códigos-fonte em Python, mas também de executá-los dentro dos notebooks. Sabe esse arquivo do qual você está lendo esse texto? Ele é um notebook, assim como os arquivos das aulas anteriores, e é por isso que é possível executar todos os códigos que se encontram nas aulas.

A extensão de um arquivo notebook é **.ipynb**, e podem ser lidos, criados, editados e executados por 3 programas que iremos listar logo mais abaixo. É possível visualizar esse tipo de arquivo diretamente do *GitHub*, mas não é possível executar os códigos diretamente de lá. Esses arquivos conseguem executar *tags* de linguagens de marcação, como o HTML, tornando possível modificar os notebooks como se fossem páginas da Internet (é possível inclusive inserir formulários HTML dentro dos notebooks). Também é possível converter os notebooks para HTML e PDF, mas em ambos os casos ele não irá mais executar os códigos em Python, e no caso do PDF, ele também não irá executar os códigos html caso eles tenham sido utilizados para criar formulários.

O que é Jupyter Notebook?

É uma aplicação web de código aberto cuja função é criar, abrir, visualizar e editar os *notebooks*. Ele funciona praticamente como uma IDE voltada para ciência mineração e análise de dados, assim como inteligência artificial. Apesar da aplicação ser web (abre diretamente do navegador padrão da máquina do usuário), ele roda localmente, ou seja, não há a necessidade de Internet para rodar o software (mas há a neccessidade de instalação).

Há 3 formas de usar o Jupyter Notebook:

- Através do ***Anaconda Navigator***: o pacote de aplicações voltadas para ciência de dados, que instala, entre outros programas, o Python e o Jupyter Notebook. Ao executar o Jupyter Notebook, ele irá abrir no navegador padrão.
- Através do **CLI**: escolher uma pasta onde irá rodar o Jupyter Notebook, abrir o terminal nessa pasta, e executar o comando `pip install notebook` (ele irá instalar o Jupyter especificamente em uma pasta). Para abrir o Jupyter Notebook, basta abrir novamente o terminal e executar o comando `jupyter notebook`, e o programa irá abrir no navegador padrão da máquina do usuário.
- Através do ***Visual Studio Code***: abrir o painel de extensões do *VSCode*, procurar pela extensão **Jupyter** e instalar. Isso fará com que os notebooks rodem diretamente do VSCode, sem a necessidade de abrir o Jupyter Notebook no navegador.

O que é Jupyter Lab?

Assim como o **Jupyter Notebook**, o **Jupyter Lab** também é uma ferramenta utilizada para trabalhar com os *notebooks*. Porém, ela é mais completa e refinada, que reúne em uma aplicação o melhor dos *notebooks* e das *IDEs* tradicionais, chegando a ser mais recomendado a utilização dele em vez do Jupyter Notebook.

Há 2 formas de usar o Jupyter Lab:

- Através do ***Anaconda Navigator***: assim como o Jupyter Notebook, o ***Jupyter Lab*** também vem dentro do pacote do Anaconda, e assim como o Jupyter Notebook, ao executá-lo, ele também irá abrir no navegador padrão da máquina do usuário.
- Através do **CLI**: é possível instalar e usar o Jupyter Lab também através do CLI. Basta abrir o terminal do Sistema Operacional na pasta desejada, e digitar o comando `pip install jupyterlab` para instalar, e após o término da instalação, executar no terminal o comando `jupyter lab` para abrir a aplicação, que também será aberta no navegador padrão do usuário.

O que é o Google Colaboratory?

Também conhecido como **Google Colab**, é o concorrente do Jupyter Notebook/Lab criado pela Google. Funciona de forma quase idêntica ao Jupyter Notebook, com a diferença de que ele não roda localmente, mas sim na nuvem. Se por um lado isso facilita a colaboração entre os cientistas/desenvolvedores em um mesmo arquivo, por

outro exige acesso constante à Internet. É necessário uma conta Google para o uso. Como era de se esperar, não há a necessidade de instalação, bastando apenas acessar o seguinte endereço logado em sua conta do Google: https://colab.research.google.com/. É possível também instalar uma extensão no seu Google Apps para que apareça a opção de criar um notebook diretamente da sua conta do Google Drive.

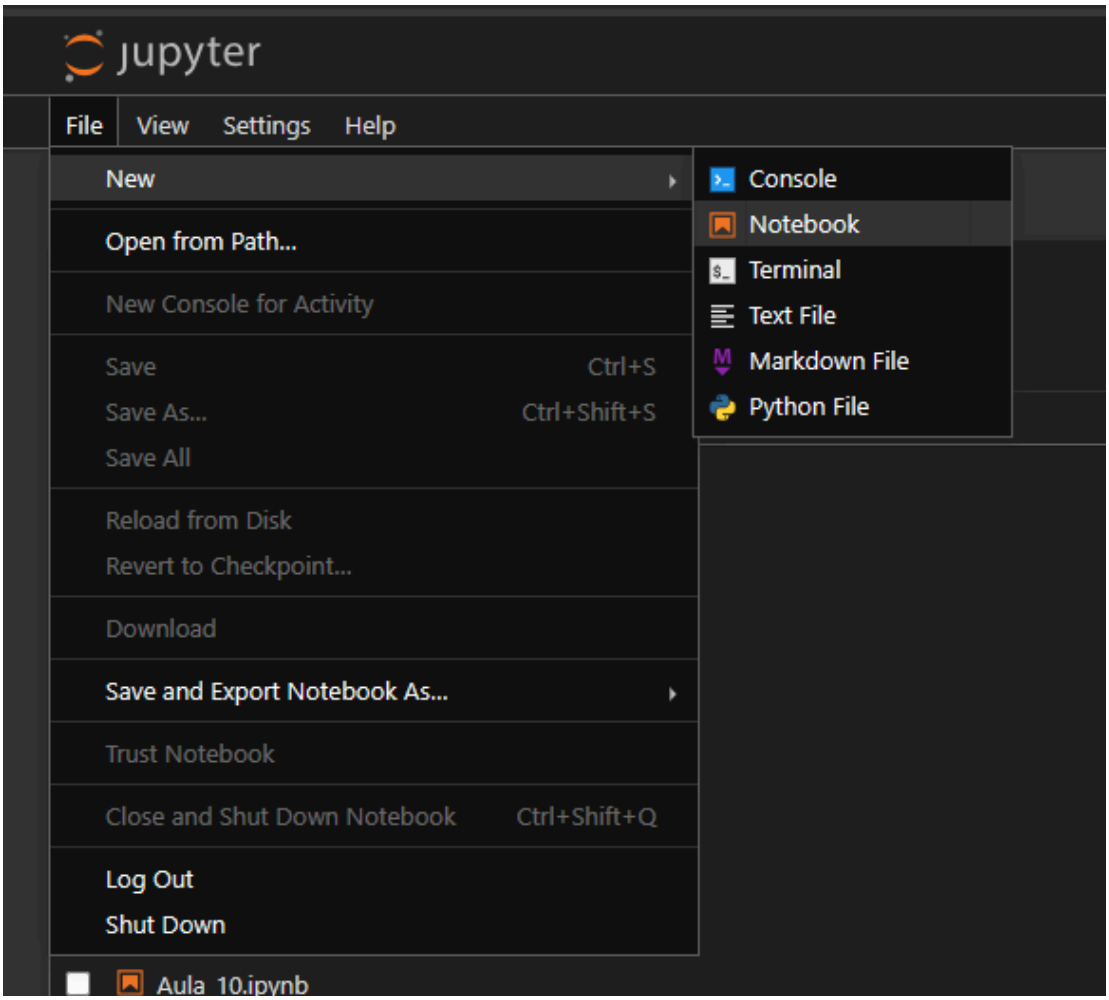
Qual usar?

A resposta simples é: tanto faz. Vai do gosto do freguês. Você usa o que se sentir mais à vontade para usar. O professor Alex se sente mais à vontade de usar a extensão **Jupyter** para o **VSCode**, pois assim é possível abrir localmente e diretamente no editor de código, sem necessidade de abrir o navegador para trabalhar com os notebooks, facilitando e agilizando o trabalho.

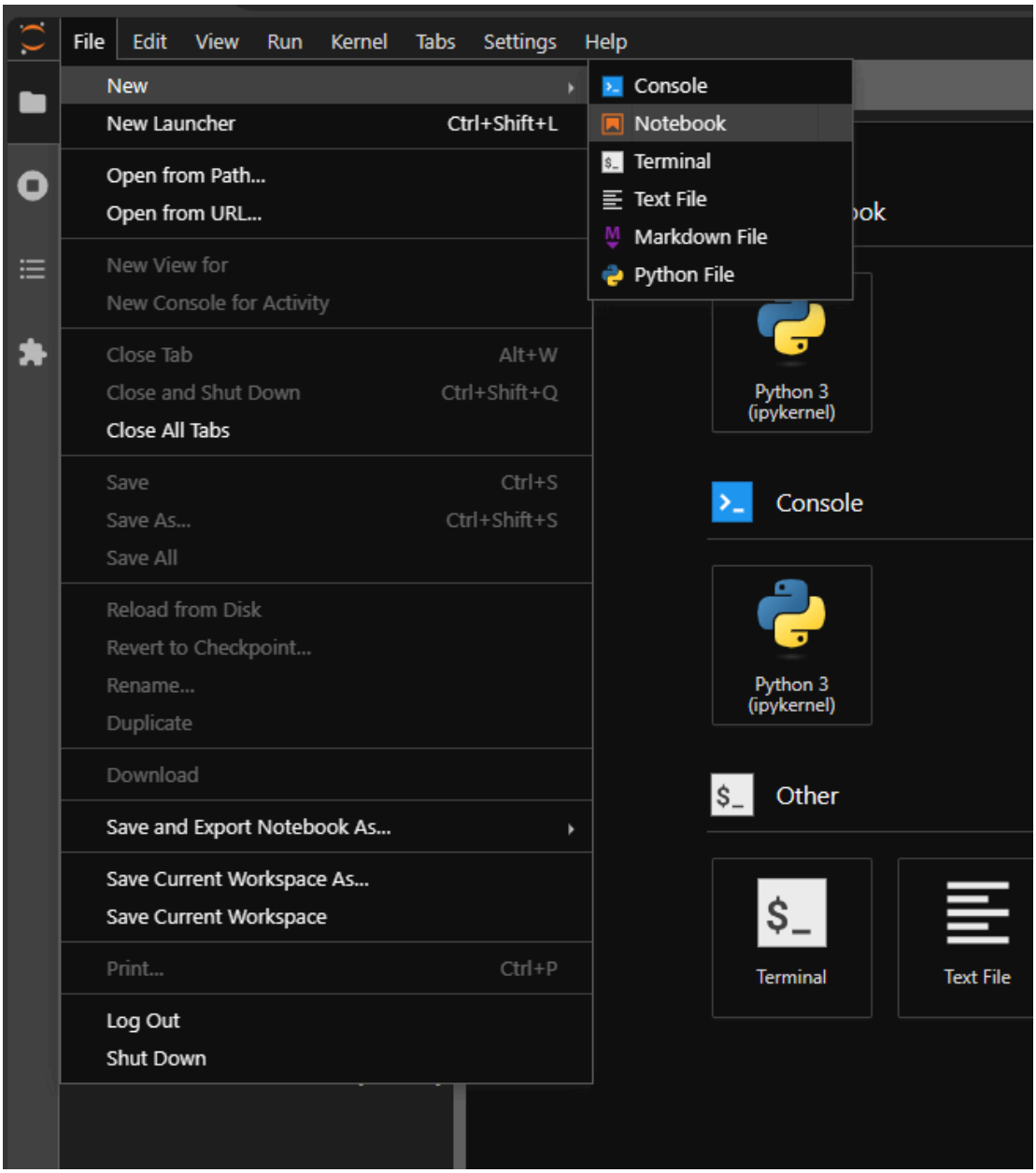
Como criar os notebooks?

Os notebooks são basicamente arquivos com a extensão **.ipynb**. Portanto, basta criar um arquivo com essa extensão que ele já será reconhecido como um notebook pelo Jupyter, VSCode ou Colab. Veja as formas de se criar um .ipynb:

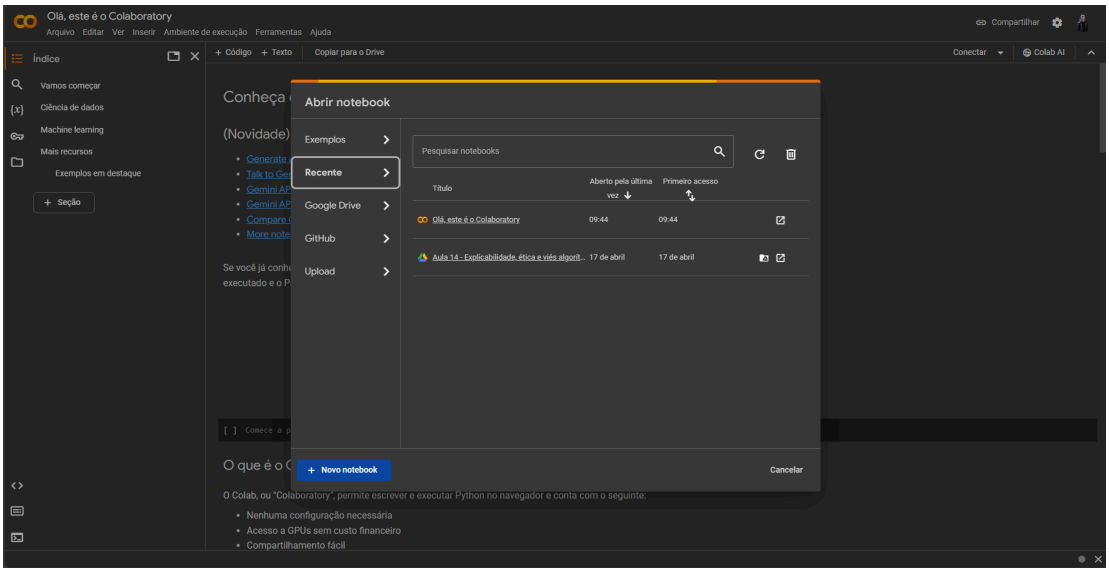
- **Criando um novo arquivo .ipynb manualmente:** qualquer forma de se criar qualquer tipo de arquivo pode ser feita para criar um ipynb. No VSCode, você pode simplesmente criar um novo arquivo no Explorador e digitar no final do nome do arquivo **.ipynb**. Exemplo: aula_15.ipynb.
- **Pelo Jupyter Notebook:** você pode abrir a aplicação **Jupyter Notebook** e ir em **File -> New -> Notebook**. Veja imagem abaixo:



- **Pelo Jupyter Lab:** você também pode usar o **Jupyter Lab**. O procedimento é o mesmo do Jupyter Notebook: **File -> New -> Notebook**. Veja abaixo:



- **Pelo Google Colab**: logado em sua conta do Google, acesse https://colab.research.google.com/, e na tela que surgir, clique no botão azul **+ Novo notebook**:



- Caso clique em Cancelar por acidente, você pode clicar em **Arquivo -> Novo notebook**:

Células do notebook

Um arquivo do tipo notebook é dividido em células. Cada célula contém um tipo de conteúdo, e podem ser de dois tipos: ***Markdown*** e ***Código***.

Markdown

Uma célula do tipo ***Markdown*** é uma célula que contém conteúdos do tipo texto e/ou imagens. A formatação do texto é feita usando alguns símbolos, por exemplo:

- `#` : título. Os níveis dos títulos mudam de acordo com a quantidade de `#` aplicados em uma linha.
- `-` : marcador de tópico.
- `---` : linha divisória.
- `_texto_` : *underscore* aplica estilo itálico quando aplicado no início e fim do texto.
- `**texto**` : dois asteríticos no início e no fim do texto deixam em negrito.
- `***texto***` : três asteríticos no início e fim de um texto aplicam negrito e itálico ao mesmo tempo.
- Acento de crase no início e fim do texto: indicam um texto que simboliza um código-fonte.

As células *markdown* também aceitam códigos *html*, o que permite uma estilização parecida com a das páginas web.

Código

Células do tipo *código* são utilizadas para escrever um código-fonte a ser executado dentro do próprio notebook. Essa é uma das principais vantagens de se usar um notebook como este. A célula exibe um código-fonte identico aos dos arquivos **.py**, e são acompanhadas por um botão de executar do lado. Ao clicar no botão, o código-fonte é executado diretamente da célula, exibindo a saída de dados logo abaixo, o que é excelente para a exibição de relatórios de análsie de dados feita por Python.

Um pequeno detalhe sobre as células de código do Notebook...

Os códigos de um notebook são executados em sequência, e não de forma isolada em cada célula. Ou seja, o valor atribuído a uma variável em uma celula permanece o mesmo caso essa mesma variável seja chamada em outra célula. Com isso, o código não precisa ser repetido caso precise ser retomado mais para frente. Esse detalhe deverá ser levado em conta a partir do próximo notebook/aula.