3.4 - Uso de bibliotecas de Python (pandas) - lendo e entendendo nossos dados

00:00:00:02 - 00:00:21:20
Olá pessoal! Na última aula nós vimos o Jupyter, na aula de hoje vamos ver os dados do módulo dois aqui no Jupyter. Bora lá. Bom, a primeira coisa que precisamos fazer aqui é acessar a nossa planilha. Aqui no Drive, a gente consegue ver a planilha na pasta que a gente criou lá para o curso e tudo mais. Bom, aqui é a planilha módulo 3.

00:00:22:15 - 00:00:48:06
Mas a gente precisa ter acesso a essa planilha lá no Colab. Então, vamos criar um novo notebook clicando no botão direito, aqui em "mais" e Google Colaboratory. Vamos criar e compartilhar. Ok, já abriu aqui um novo notebook. Vamos colocar uma célula de texto aqui no início pra dar um nome dessa aula e deixar já tudo organizadinho. Vamos lá!

00:00:48:09 - 00:01:10:04
Uma célula de texto. Ela vai aparecer embaixo dessa célula de código que já começa. A gente pode usar essas setinhas aqui. Tem pra cima e para baixo. A gente coloca pra cima, pra poder ficar em cima. Uma coisa interessante aqui é que a gente pode colocar o jogo da velha pra mudar o tamanho da letra do título.

00:01:10:23 - 00:01:40:21
Eu coloquei um jogo da velha só e vou colocar aqui o nome da aula de Uso da Biblioteca Pandas. Uso da Biblioteca Pandas. Executo a célula e está aqui bem grandão. Ah, só lembrando: se você colocar mais jogo da velha aqui, o tamanho da letra vai diminuir. Existem dois modos de a gente acessar o drive aqui no Colab.

00:01:41:05 - 00:02:14:27
O primeiro é a gente vir nesse canto esquerdo, clicar nessa pastinha de arquivos e vai aparecer aqui toda essa lateral de arquivos. Se a gente clicar nessa pasta que tem esse simbolozinho do drive, ele vai montar o drive. O que significa montar o drive? É criar um ponto de acesso ao drive. A gente pode vir aqui em conectar ao Google Drive. Ele vai estar montando o drive e aqui já aparece a nossa pasta de drive.

00:02:15:10 - 00:02:38:20
Se a gente clicar nessa setinha, vai aparecer outras pastas. A gente tem aqui o nosso "My drive" e aqui a gente já tem o nosso ponto de acesso ao nosso drive. Uma segunda forma é importando o drive direto em uma célula do Colab. E aí, primeiro a gente importa o Colab. Então: from google.colab. Ou seja, do Google Colab "import drive".

00:02:38:20 - 00:03:13:18
Importamos o Drive e vamos executar mais uma célula de código. E aí agora a gente vai criar o ponto de acesso para o drive, que é drive.mount, de montar o drive, como tinha aparecido aquela mensagem no cantinho. E aí a gente coloca qual é o caminho que a gente quer acessar. Aqui a gente coloca content drive. E aí a gente executa essa célula.

00:03:13:18 - 00:03:42:20
E aí vai aparecer se a gente quer realmente permitir que o notebook tenha acesso aos arquivos no drive, ou seja, a gente está realmente criando esse ponto de acesso ao Drive. A gente coloca "conectar ao Google Drive". Seleciono qual que é a conta, vou continuar, continuar de novo e pronto. Já vai falar aqui: olha, a gente criou esse ponto de acesso no Drive lá na pasta Drive.

00:03:43:12 - 00:04:13:22
E aí se não tivesse aparecido, se a gente não tivesse já clicado por aqui e já acessado o drive, aí iria aparecer a pastinha de drive aqui no canto. Em uma aula anterior, nós discutimos rapidamente o que são bibliotecas, que fazemos a importação das bibliotecas e utilizamos as funções que precisamos. Nessa sala, nós vamos utilizar a Biblioteca Pandas, que é amplamente utilizada na área de análise de dados, pois é uma ferramenta que fornece várias funções para análise e manipulações de dados.

00:04:14:04 - 00:04:43:28
Então vamos criar uma célula de código e importar no Pandas. Ó, mais uma célula de códigos: import pandas. Mas para facilitar, nós vamos dar um apelido para o pandas. Assim, a gente não precisa ficar digitando pandas toda vez que a gente for usar biblioteca. Para isso, nós escrevemos assim: import pandas as pd. Esse "as" é para dizer que vamos chamar o pandas por um apelido.

00:04:43:28 - 00:05:09:26
É algo como importar o pandas com o nome de "pd". Assim, em vez de digitar pandas, vamos digitar apenas "pd". Quatro letras a menos, já ajuda na digitação. Então, a primeira função do pandas que a gente vai utilizar é para ler o nosso arquivo que está lá no drive. O formato do nosso arquivo é em Excel. Então, no pandas nós temos uma função que é read excel, ou seja, ler o excel.

00:05:10:10 - 00:05:39:14
A gente chama essa função assim... Primeiro vamos executar aqui o import do pandas, mais uma célula de código. E aí é: pd.read_excel. E aí nesses parênteses, aqui dentro, que a gente vai colocar o caminho que está o nosso arquivo. Bom, pra encontrar o arquivo, primeiro a gente clica aqui no drive, dentro do drive tem o "my drive", que é o nosso drive.

00:05:40:12 - 00:06:05:12
E aí aqui dentro a gente vai procurar qual é a pasta que a gente criou, que a gente salvou o nosso arquivo. No meu caso, eu criei essa pasta aqui de PrograMaria, pra ficar bem organizadinho. E aí a gente procura dentro dessa pasta qual é o arquivo. Aqui é planilha módulo 3. A gente clica nesses três pontinhos aqui e clica em Copiar Caminho. E isso já vai ter copiado o nosso caminho.

00:06:05:13 - 00:06:28:06
A gente pode vir aqui dentro dos parênteses agora e colocar ctrl+v pra colocar o caminho. Se a gente executar, vai dar um erro porque a sintaxe, o modo que a gente colocou está errado. O que a gente precisa fazer? A gente precisa colocar todo esse caminho entre aspas. Pode ser aspas duplas ou aspas simples.

00:06:28:23 - 00:06:39:16
Então vamos fazer um ctrl+z só pra poder retirar, colocar aspas e agora sim ctrl+v para colocar o caminho aqui dentro. E aí a gente executa a célula.

00:06:42:23 - 00:07:08:06
Bom, aqui embaixo já vai aparecer a planilha pra gente. Agora a gente pode até fechar esse cantinho aqui que a gente já copiou nosso caminho do arquivo. Aqui a gente tem a nossa tabela, só que a gente precisa colocar essa tabela numa variável para a gente poder conseguir acessar essa tabela mais. Então eu vou chamar essa variável de “dados” aqui.

00:07:08:27 - 00:07:40:00
Agora vou executar essa linha. Não apareceu a tabela mais. Você pode pensar: Ai meu Deus, sumiu a tabela. Não, agora a gente atribuiu essa tabela a essa variável dados. Então, toda vez que eu chamar “dados”, vai ter a nossa tabela. Se eu clicar aqui em dados, é só executar dados e já vai aparecer a tabela. Reparem que, diferente do Excel ou Google Planilhas que podíamos só arrastar a tabela e a gente conseguia ver todos os dados, aqui não aparece tudo, aparece algumas linhas iniciais e algumas finais aqui.

00:07:40:11 - 00:08:05:28
Algumas linhas do início, que a gente consegue ver aqui: 1, 2, 3 e 4, e lá para o final da planilha. Isso porque se aparecesse tudo nosso notebook ia ficar gigantesco. Nós temos uma função muito legal para ver as planilhas, as linhas iniciais da tabela, que é o head, de cabeça em inglês.

00:08:06:11 - 00:08:40:16
Se em uma próxima célula a gente digitar, por exemplo, mais uma linha de código aqui, dados (que é a nossa tabela) ponto head e executarmos, vai aparecer as cinco primeiras linhas da nossa tabela. É um padrão dessa função head aparecer as cinco primeiras, mas caso a gente queira as dez primeiras linhas, basta a gente colocar o número aqui dentro desses parênteses.

00:08:40:16 - 00:09:09:26
Então se eu vier aqui, colocar dez e executar, vai aparecer então as dez primeiras linhas da nossa tabela. Temos uma função para ver as linhas finais da planilha também. Na próxima célula aqui, mais uma célula de código, a gente coloca dados.tail e os parênteses, aí a gente executa. E aqui a gente vai ter as últimas linhas da nossa tabela.

00:09:09:26 - 00:09:32:20
Também é por padrão aparecer as últimas cinco linhas mas se a gente quiser um outro número de últimas linhas, a gente coloca aqui entre os parênteses da função tail. Uma coisa interessante é que tem aparecido uma ordem de números aqui no canto, no canto esquerdo da nossa tabela. Esse é o nosso índice, ou index. Por padrão é assim, mas podemos colocar qualquer valor, até texto, e não necessariamente precisa ser ordenado.

00:09:33:04 - 00:09:58:13
Mas lembrem de prestar atenção nele, porque vai ser importante em outras funções mais pra frente. Bom, mas por padrão, o index vai ser um número das linhas em ordem. As contagens em Python sempre começa do zero. Por isso com o head aqui em cima, a gente vai ver que começa com o zero: 0, 1, 2, 3. E com o tail veremos o index das últimas linhas.

00:09:58:13 - 00:10:37:19
Por isso a gente tem aqui quatro mil duzentos e tal. Através da função shape, a gente consegue ver o tamanho da nossa planilha. Vamos colocar aqui mais uma célula de código: dados.shape. O shape não precisa dos parênteses, porque ele é um atributo da tabela e não uma função. Lembra da nossa função de lista de compras? O head é como uma função porque ele precisa de um algoritmo pra selecionar cinco ou dez primeiras linhas. Já o shape é como um preço cenoura, só um atributo com um valor, ele apenas retorna o tamanho da planilha em linhas e colunas.

00:10:37:19 - 00:11:13:11
Então o primeiro valor aqui vai ser as linhas e o segundo as colunas. Portanto, podemos dizer que a nova tabela tem 4271 linhas e 28 colunas. Tudo certo? Uma outra forma de saber quantas linhas tem a nossa tabela é usando a função len. Ficando assim... Vamos criar aqui mais uma célula de código. Len, parênteses e dentro desses parênteses a gente coloca nossa tabela. Executamos e aí a gente tem o valor de 4271.

00:11:13:21 - 00:11:42:21
Para saber quais são as colunas da tabela, vamos pegar o atributo de columns, coluna em inglês. Para isso, vamos escrever e executar em uma nova célula: dados.columms e aí a gente executa. E aqui tem todos os nomes de todas as colunas. A gente tem mais uma função, que é a função info. Então vamos colocar aqui mais uma célula de código: dados.info. Vamos executar.

00:11:43:00 - 00:12:06:14
E olha que legal essa saída a gente tem aqui o range do index, que aqui mostra pra gente a quantidade de entradas, ou seja, a quantidade de linhas que a nossa tabela tem e o range que vai de 0 a 4270. A gente tem também o total de colunas, com 28 colunas. A gente tem aqui também os nomes das colunas e os indexes delas aqui do lado e o nome delas.

00:12:07:01 - 00:12:32:21
Então, aqui embaixo a gente tem a memória utilizada, a quantidade de memória utilizada por essa tabela e a gente tem aqui os tipos de dados de cada coluna. Ou seja, a gente tem booleano, float. Mas a gente tem esse dado aqui também: se a gente arrastar para o lado aqui, a gente tem a quantidade de linhas não nulas, ou seja, a quantidade de linhas preenchidas em cada coluna e a gente tem o tipo.

00:12:34:10 - 00:13:05:17
Vamos fazer um retrospectiva rápida aqui: os dados podem ser do tipo qualitativo, referente a qualidade, como por exemplo, cor, gênero, etc. Podem ser do tipo quantitativo: valores que são mensuráveis, como por exemplo, idade, preço, etc. E também temos os dados categóricos, responsáveis por categorizar algo. Exemplo o nível de escolaridade. Temos os dados discretos também, que são basicamente números inteiros e não negativos de zero a infinito, mas apenas números inteiros.

00:13:05:27 - 00:13:31:20
Exemplo a quantidade de pessoas em uma sala. E os dados podem ser do tipo contínuos, que são dados numéricos porém os números podem ser negativos quebrados, ponto flutuante ou float, como vimos aqui no nosso arquivo, e podem ser de qualquer valor, tipo altura, que é 1,64, que é um número quebrado. Relembrar os tipos de dados é importante para a gente olhar uma coluna e já saber qual tipo de dados temos aqui.

00:13:32:03 - 00:13:50:14
Então, se a gente olhar aqui nesse tipo de dados, a gente tem o int64. A gente sabe que é um número inteiro. A gente tem um dado tipo objeto, então a gente sabe que é um tipo string, que é texto. A gente tem o float que a gente sabe que é um dado de tipo contínuo, que pode ter números quebrados, ponto flutuante.

00:13:50:25 - 00:14:23:23
A gente tem uma tabela aqui, que é do tipo bool, ou seja, booleano, que aceita verdadeiro ou falso, true ou false. Enfim, a gente tem bastante coluna do tipo objeto, mas fiquem tranquilos que a gente vai dar uma olhadinha mais pra frente. Bom, uma última função aqui muito interessante... Vamos colocar aqui, mas uma célula de códigos: dados.describe, ou seja, descrição em inglês. A gente executa e aqui a gente tem mais uma tabela.

00:14:24:01 - 00:14:49:17
Aqui não apareceu todas as colunas. Por que? No describe, ele vai dar o retorno para a gente das colunas, que tem valores numéricos, ou seja, a gente tem idade, mudou de estado, gestor, salário e essa unnamed, que é uma coluna que só quando a gente salva um arquivo pode ser que tenha essa coluna, mas ela não interessa pra gente nesse momento.

00:14:50:19 - 00:15:16:10
Então no describe ele vai retornar essas colunas que têm valores numéricos e aqui nas linhas a gente tem algumas informações interessantes, como a contagem de linhas com entradas, ou seja, colunas que têm algum valor. A gente tem a média, desvio padrão, mínimo, máximo, alguns percentuais aqui de quartis. Não se preocupem que a gente vai também ver em uma próxima aula sobre o que é isso de média, desvio padrão e tudo mais.

00:15:16:10 - 00:15:43:14
Mas o describe é muito interessante pra gente poder dar uma visualizada geral em colunas com dados numéricos. Nessa aula nós vimos várias funções da Biblioteca Pandas, como por exemplo o head, tail, info, describe, shape. E eu quero deixar uma dica aqui muito legal que é o cookbook do Pandas. Vamos abrir uma nova guia, nova aba, na verdade, e digitar aqui ou na barra de pesquisa aqui do Google mesmo: cookbook pandas.

00:15:45:09 - 00:16:08:14
E aí, nessa primeira opção que aparecer aqui, a gente já pode clicar. O cookbook é basicamente um livro de receitas. No caso do Pandas, essa página tem várias dicas de utilização de diferentes funções dessa biblioteca. Portanto, ficou na dúvida de alguma função, procure aqui na documentação oficial da biblioteca para entender melhor como funciona a função que está tentando utilizar.

00:16:09:07 - 00:16:31:22
Infelizmente, a página é toda em inglês, mas podemos usar a tecnologia a nosso favor e copiar o que está aqui para o Google Tradutor, por exemplo, ou traduzir a página para português. O que não pode é ficar com dúvida. Outra coisa é que a gente não necessariamente precisa fazer a análise de dados em Python, mas podemos perceber reconstruindo a análise em um notebook que temos uma história ali sendo contada.

00:16:32:04 - 00:17:01:24
É fácil pra outra pessoa entender o que estamos fazendo e como chegamos em respostas para as nossas perguntas. Além disso, o Python facilita na leitura, visualização e na lógica de programação do código, podendo ser utilizada em várias etapas de um projeto de análise ou ciência de dados. Fora que para grandes volumes de dados não tem como fazer no Excel. Dependendo do volume, às vezes é até impossível, mas fica super lento e em Python tem mais opções de análise e automatizações que o Excel não permite, beleza?

00:17:02:11 - 00:17:04:07
E é isso por hoje, pessoal, até a próxima.