06

Dataset do projeto

Transcrição

[0:00] Galera, maravilha, vamos iniciar nosso curso de estatística, começar a colocar a mão na massa. Estou aqui com o colab aberto, a gente vai, como vimos nas aulas anteriores, entrar no endereço colab.research.google.com, a gente vai ter que se logar pra fazer upload dos arquivos e acompanhar as aulas, faz o login aí, estou com a ferramenta aberta, eu deixei pra vocês nos recursos associados a essa aula dois arquivos pra fazer download, então são esses aqui, curso de estatística parte 1, .ipynb, que é notebook que eu preparei pra gente acompanhar as aulas, pra você manter como material de estudo, está tudo documentadinho, a gente vai só preencher as células vazias. Tem aqui também um dataset que é o dados.csv, que é o dataset que a gente vai utilizar na aula, vamos falar sobre ele daqui a pouquinho.

[0:54] Então, vamos fazer inicialmente, abrir, fazer o upload do nosso notebook, vem em file, upload notebook, vai abrir essa janelinha, vem na última aba que está marcada, upload, escolher arquivo, navega até a pasta onde você salvou o notebook da nossa aula, e abre aqui, clica em abrir, vai demorar um pouco, depende da sua conexão, mas vai abrir o notebook.

[1:27] vou entrar em envio colab sections, pra fechar, pra mostrar pra vocês aqui tem o table of contents, faz um sumário de tudo que tem no nosso notebook, você pode acompanhar aqui, clicar até a parte que você deseja, que você quiser.

[1:42] mas eu fechei pra gente ver os tópicos principais que vamos ver na parte 1 do curso de estatística, mais um curso de estatística descritiva, que é a primeira fase de um processo de análise, quando a gente começa a estudar os dados que estamos trabalhando.

[1:57] Conhecendo os dados, vou mostrar pra vocês o dataset que vamos utilizar, depois distribuição de frequência, medidas de tendência central, médias, mediana, moda, coisas que você deve ter ouvido falar, as medidas separatrizes, que são, incluem a mediana, são aquelas quartil, decil, a gente vai estudar todas elas, e as medidas de dispersão, variância, desvio padrão, que você também já deve ter ouvido falar. Aqui a gente vai entender melhor como eles funcionam, quando utilizar, como utilizar, como calcular, e vamos, pra isso, utilizar a biblioteca Pandas do Python.

[2:39] Vou aqui abrir todo aqui, o contrário que a gente fez, eu fechei, agora vou expandir as sessões, expandiu, o que eu quero mostrar pra vocês, juntamente vamos fazer o upload do arquivo, vamos primeiro fazer o upload, porque caso demore, ele já vai fazendo.

[2:54] Eu entro aqui na guia Files, está vendo, Files.

[3:01] Vou esperar que está conectando, conectou, tem o botão escrito upload, e venho aqui, pego o dado que a gente fez o download e abro ele. Ele vai vim com a sua lembrança aqui, você clica em ok, está dizendo que quando você fechar aqui, isso tudo vai ficar perdido, então toda vez que a gente começar, você fechar sua sessão, a gente for fazer novamente, você tem que fazer todo esse processo de novo, fazer o upload do arquivo, abrir o notebook, no final da aula vou mostrar como faz o download de um notebook, com as modificações que a gente fez, então, toda aula você vai ter que fazer isso aqui, infelizmente, no colab funciona desse jeito.

[3:41] Então, vamos lá, esse dado que abri agora é um dado que busquei no site do IBGE, instituto brasileiro de geografia e estatística, instituto oficial de estatística do Brasil, está aqui um link pra ir diretamente onde peguei os microdados, e o que são microdados?

[4:02] no caso da PNAD, a pesquisa que escolhi pra utilizar o dataset, é como se você estivesse vendo a entrevista, cada registro de um microdado é a entrevista que o recenseador, o entrevistador foi na casa da pessoa e fez, ou num

estabelecimento comercial, por aí vai. Aí você tem essa informação, lógico, não está identificado, não sei quem foi que respondeu aquilo, mas você tem informação do questionário.

[4:30] É a coisa mais agregada que você consegue de um dataset de uma pesquisa estatística, aqui estamos utilizando PNAD, está no ano de 2015 porque depois do ano de 2015 fizeram uma modificação, mas isso não vem ao caso, se você quiser pesquisar, entra no site do IBGE e dá uma procurada do que se trata PNAD, das modificações que ocorreram. Aqui tem, no notebook, uma descrição básica da PNAD que tirei do próprio site, você pode dar uma lida pra entender melhor.

[5:05] eu tratei esse dataset utilizando algumas informações que disponibiliza pra gente fazer um download, leitura em R, você deve ter ouvido falar em R, um software estatístico, tem também a leitura dos dados em insight, que também é outro pacote estatístico, então você tendo curiosidade, você entra ali, faz o download, faz os trabalhos que você quiser com esses dados, estou mostrando o que eu fiz pra gente acompanhar nossa aula. Aqui eu peguei uma variável de renda mensal do trabalho principal pra pessoas de 10 anos ou mais de idade, uma variável que eu escolhi pra rendimento, outra variável é a idade do morador em anos, essa variável de altura, eu criei essa variável, elaborei aqui, depois posso deixar uma dica pra você dar uma olhada em como criei essa variável, pra não ficar uma coisa mágica que apareceu aqui, mas eu criei por um motivo que eu vou falar nos próximos vídeos. A altura do morador é uma coisa aleatória, foi criada por mim, unidade da federação, os estados do Brasil, estão aqui.

[6:15] vou mostrar que nosso Dataset não está com Rondônia, Acre, Amazonas, não está com descrição, está com um código, que se você quiser associar, é só vir nessa tabela que seria nosso dicionário. Sexo também, 0 pra homem e 1 pra mulher, anos de estudo aqui tem uma codificação que está no nosso dataset, codificado, e aqui a descrição, que significa o código, e aqui, cor ou raça, não botei raça mas é a mesma coisa, indígena, branca, preta, amarela, parda, são as cores e raças das pessoas.

[6:51] aqui uma coisa importante, você que é estatístico, estudante de data Science, cientista de dados, sempre lembrar de documentar o que você fez no dataset pra poder lembrar futuramente quando for utilizar seus dados.

[7:07] Realizei aqui o seguinte tratamento nesses dados, pra poder utilizar na nossa aula, eliminei os registros de renda que eram inválidos, na PNAD eles vêm com esse código, 12 números 9, eliminei também os registros de renda que eram "Missing", ou seja, não tinham renda nenhuma, nem 0 nem nada, cortei fora, as que tem 0 ficou, as que tem "missing", não. Missing é um valor nulo. Cortei fora, e considerei somente os registros da pessoa de referência, que é a pessoa que responde o questionário, como se fosse o chefe da família, esses são os 3 tratamentos que fiz nesse dado.

[7:43] vamos fazer o seguinte, iniciar com o Pandas e fazer a leitura dos dados, está com CSView, quero passar pra uma estrutura que é o Data Frame, você deve conhecer, se você não conhecer o Pandas, é um pré-requisito desse curso saber um pouco de Pandas, mas eu vou andar devagar com o Pandas, não vou fazer nada muito mirabolante sem explicar o que estou fazendo. Vamos primeiro importar o Pandas, Pandas SPD, que é o apelidinho que a gente dá pra ficar mais fácil de digitar, está importando, às vezes o colab demora um pouquinho, porque você entra numa fila pra rodar as requisições. Eu vou atribuir todo aquele dataset, uma variável que vamos utilizar, vou chamar de Dados, pra isso vou chamar o Pandas PD, Read CSV, estou lendo o arquivo CSV que é o que fizemos o upload aqui, está no nosso dado, dados.csv.

[8:47] ok, e vou passar pra aqui, tá aqui do lado, só precisa fazer isso, dados CSV. Tá bom?

[8:54] Eu já li esse arquivo e transformei num dataframe, só pra mostrar pra vocês, vamos abrir isso aqui, type dados, um Pandas com Dataframe, um dataframe do Pandas, posso visualizar ele utilizando o Colab, uma ferramenta bem próxima do Jupiter, shift enter, ou então você clica aqui mostrando, você tá vendo, seleciona aqui, tem uma setinha de play, ele roda, ou aperta shift-enter, ele vai rodar também.

[9:36] Então, aqui tem um conjunto de dados, que eu separei pra gente, tem UF, Sexo, Idade, cor, anos de estudo, renda, e altura. Próximo passo é a gente tentar entender como são constituídas essas variáveis, como a gente classifica essas variáveis, porque isso é importante, algumas variáveis classificadas de algumas formas, são tratadas de formas diferentes também, mas isso a gente vai ver no próximo vídeo, vejo você lá.