03

Conhecendo o ambiente e o dataset

Transcrição

[0:00] Legal, pessoal, vamos começar agora nosso curso, continuação do nosso curso de estatística, estamos na parte 2 agora.

[0:08] Inicialmente, vou mostrar para vocês a ferramenta que vamos utilizar para desenvolver o nosso treinamento, que é o Colab, que foi a mesma ferramenta que utilizamos no parte 1.

[0:18] Então, você entra aqui no Google, digita Colaboratory, com Y no final, e a primeira opção que vai aparecer, você clica, precisa estar logado para fazer upload dos arquivos, do mesmo jeito que a gente fez no curso parte 1. Eu vou abrir inicialmente, vou vir em File, upload notebook, e escolher o arquivo, vou vir aqui em Versões, bibliotecas.

[0:39] Vou abrir esse aqui, deixei esse para você fazer o download, verificar as versões que você tem, ele rodou aqui, eu tinha rodado isso, aqui estou utilizando o Pandas, Numpy, Scipy, e matplotlib, então, agora, estou utilizando essas versões, você que está no futuro, provavelmente pode vir outro tipo de versão, se você tiver algum problema enquanto estiver executando o código, você pode voltar, fazer o downgrade aqui no colab, a gente ensinou no curso 1 de estatística, se você tiver dúvida, pode ver lá no primeiro vídeo.

[1:11] Isso que eu queria mostrar. Vou vir aqui agora, abrir a aba lateral, vou primeiro fazer a carga do arquivo, dos dados, é outra coisa que quero mostrar para você, os dados que quero utilizar para realizar o nosso projeto, é o mesmo dataset, vem aqui, dados, CSV, está aí para fazer download também, o mesmo dataset que utilizamos na parte 1, vou utilizar nesse também, nesse treinamento não vamos utilizar muito esse dataset, mas no final eu vou deixar como fiz no primeiro, um notebook para você executar um projetinho usando o que aprendemos nesse treinamento.

[1:52] Vamos falar de probabilidade, estimação, então vou deixar um notebook com alguns exercícios, legal? Carregou os dados, está aqui, Dados, CSV, está carregado, vou abrir agora o Notebook que preparei, File, Upload Notebook, escolher arquivo, curso de estatística, parte 2. Também está aí, nos recursos, para você fazer download. Preparei esse notebook para gente seguir um roteirinho para executar nossa aula aqui. Então está aqui, primeira parte, quero mostrar justamente isso, a gente conheceu o dataset, o mesmo do curso anterior, é um dataset que peguei do site do IBGE, ele vem da pesquisa nacional por amostra de domicílios de 2015.

[2:33] Então, aqui, a fonte dos dados, as variáveis que vamos utilizar, dentro do nosso dataset, renda, idade das pessoas, altura foi uma elaboração minha para fins didáticos, para estudarmos uma distribuição que se comporta de forma normal, na distribuição de programa, normal é uma coisa que vamos ver nesse treinamento agora.

[2:51] Aqui tem unidade da federação, os códigos da unidade da federação, não está escrito os nomes dela, tem os códigos. A gente tem sexo também, código, masculino e feminino, anos de estudo, tudo codificado, cor ou raça, aqui também.

[3:09] E aqui algumas observações de tratamentos que realizei no dataset, quis realizar para facilitar nosso aprendizado, único e exclusivamente para isso. Eliminei os registros de renda que são inválidos, tem um código aqui, um monte de 9. Cortei isso fora. Cortei a renda, que é Missing também, não tem renda no registro da pessoa, cortei fora também.

[3:34] E, eu considerei somente os chefes de domicílios, pessoas de referência, as pessoas que foram entrevistadas, cortei isso também, é importante, sempre que vai fazer um tratamento de um dado, escrever esse roteirinho aqui para gente saber o que aconteceu, por que os resultados que estamos obtendo estão daquele jeito, talvez por causa de um tratamento que a gente realizou, ok?

[3:57] Então, vamos importar este dataset, eu trouxe aqui pro Colab, dentro do arquivo, esperar um pouco. Está vindo, veio. Dados CSV. Então, vou usar o Pandas como eu disse, Import, Pandas SPD, aquele apelidinho que a gente já está acostumado.

[4:23] Vou ler esse arquivo, colocar dentro da variável "Dados", então chamo Pandas.read.csv, e passo o dados aqui, .csv. Esse carinha que está aqui.

[4:39] Shift-enter, ele já rodou, então vou visualizar os primeiros cinco registros, como que faço isso? Dados.read, e eu consigo visualizar os cinco primeiros registros, para gente ter uma ideia do que a gente tem dentro desse dataset.

[4:55] Está aqui a UF, como diz, tudo em código, sexo em código, idade, são números, lógico, cor também está em código, anos de estudo, também codificado, aqui a renda, está em reais, e a altura, em metros, como eu disse, a altura foi uma elaboração minha, é uma variável fake.

[5:15] Então, já conhecemos a ferramenta, conhecemos o dataset que vamos utilizar um pouco nesse curso, pouca coisa, eu vou tentar utilizar mais no projeto final, e a partir de agora começamos, vamos colocar a mão na massa e falar das distribuições teóricas de probabilidade, no próximo vídeo a gente vê isso, abraço.