03

Ambiente e conhecendo o dataset

Transcrição

[0:00] Pessoal, maravilha. Antes de começar o nosso treinamento, vamos dar uma olhada no ambiente que a gente vai utilizar para codificar e também nos dados que a gente vai estar utilizando. Legal?

[0:09] Eu vou utilizar, como fiz nos outros cursos, o Colab. Digitando aqui no Google, Colaboratoy, com Y no final. A primeira opção é ele, então você clica aqui.

[0:18] O objetivo aqui é a gente verificar as versões que eu estou utilizando agora, quando eu estou gravando, para você quando estiver fazendo o curso aí no futuro.

[0:27] Se tiver algum probleminha na execução de algum código, você dá uma olhada na versão que você está usando; se for diferente da minha, pode ser esse problema - você pode voltar uma versão ou duas, sem menor problema.

[0:36] No Curso Um, no vídeo de apresentação eu mostro como fazer essa volta de versão. Legal?

[0:42] Então vou dar um Esc aqui porque ele já abriu uma janela. Eu deixei dois arquivos para você fazer download, dois Notebooks. Um é o do nosso curso, que a gente vai executar o curso, e o outro é para a gente verificar as versões.

[0:55] Vamos lá. File. Dou Upload no Notebook, ele vai vir aqui na aba Upload direto, e escolher arquivo. Versão Bibliotecas, abre esse carinha aqui. Ele já vem com um macetinho aqui para a gente verificar as bibliotecas que a gente está utilizando.

[1:11] Vou rodar aqui, é só clicar Shift + Enter. Ele vai rodar de novo, o meu já tinha rodado. Nós vamos fazer novamente. Está executando, demora um pouquinho. Ok.

[1:20] Versão do Pandas, essas são as bibliotecas que eu vou utilizar. Pandas 0.24.2. Numpy 1.16.3. Scipy 1.2.1. Statsmodels 0.9.0.

[1:37] As bibliotecas últimas aqui, de visualização de gráficos, a gente vai ver uma coisinha ou outra, é pouca coisa. O Seaborn 0.9.0 e o Matplotlib 3.0.3. Perfeito?

[1:51] Executa aí, vê qual é a versão que você está usando e qualquer coisa você volta uma versão.

[1:55] Fechando aqui, eu já deixei aberto o Notebook que a gente vai começar a trabalhar. Ele já está todo organizadinho, a gente já vai começar a falar de teste de hipótese aqui nesse curso, legal.

[2:06] Mas a primeira coisa que eu quero mostrar é o dado que a gente vai vez ou outra utilizar no nosso treinamento. Tá bom? E depois, no final, tem sempre um Notebook lá, com os exercícios, utilizar esse Dataset para ficar um projeto bem legal.

[2:20] Então, vamos lá. Abrindo aqui, conhecendo os dados, o Dataset é o mesmo que a gente vem utilizando nos outros cursos. É um dataset que eu obtive da pesquisa nacional por amostra de homicídios do IBGE, no ano de 2015. Ele tem as seguintes variáveis: renda; idade; altura foi uma variável que eu fiz, eu elaborei ela a partir de uma variável aleatória normal para fins didáticos.

[2:43] Para a gente poder fazer uns exemplinhos, entender o funcionamento de algumas coisas; UF, que é unidade da federação; sexo; anos de estudo; e cor, que é a raça. Todas no Dataset estão codificadas, estão em número. Se você quiser transformar na descrição é só executar um procedimentozinho simples.

[3:04] Observações importantes. É bom a gente isso sempre que a gente estiver trabalhando com Dataset. Eu fiz alguns tratamentos, estão aqui. Foram eliminados registro de renda quando eram inválidos, quando era missing.

[3:17] Eu só assumi as pessoas de referência do domicílio, ou seja, eu só estou pegando as pessoas de referência. Talvez sejam os chefes do domicílio, eu estou chamando de chefes do domicílio. Perfeito.

[3:29] Vamos abrir agora esse nosso Dataset. Primeira coisa, importar as bibliotecas que a gente vai utilizar. Eu vou usar o sistema de ir importando bibliotecas conforme a gente for precisando. Mas, inicialmente a gente já vai precisar do Pandas pelo menos.

[3:43] Então vamos lá. Import Pandas as PD. E eu também, para economizar, a gente vai utilizar o Numpy em alguns pontos do nosso treinamento, então eu já vou importar o Numpy também, as NP. Com esses dois apelidos que são comuns na comunidade. Panda as PD, Numpy as NP. Ok? Shift + Enter. Rodou, Ok.

[4:05] Vamos agora criar um Dataframe. Os dados, eu esqueci de mostrar. Eu vou mostrar aqui. Abrindo essa aba lateral, vindo aqui em Files - nos outros cursos a gente fez isso e o meu já está aberto aqui, por isso que eu esqueci.

[4:18] Eu posso vir aqui em Upload. Os dados já estão aí para você fazer download. Aqui, Dados, você clica nele. Ele vai subir o arquivo que você vai poder utilizar ele.

[4:29] Lembrando também que você precisa estar logado no Colab para poder fazer esse Upload, para poder executar o nosso curso numa boa.

[4:36] Então, eu estou logado - aqui uma mascarazinha de mergulho. Ok. Fechar aqui para aumentar a nossa tela. E vamos lá. Criar uma variável Dados e vou chamar o Pandas.

[4:49] A função Read.csv, para ler o arquivo CSV e colocar ele dentro de um Dataframe, e vou passar aqui Dados.csv, que é o arquivo que a gente fez o Upload lá.

[5:01] Tudo certinho. Shift + Enter. Rodou. Vou dar uma visualizada nos cinco primeiros registros. Dados.head. Shift + Enter. Está aqui, o Dataset já está carregado com UF, sexo, idade, cor, estudo, renda e altura.

[5:19] Aqui algumas, como eu disse, estão codificadas em número. Por exemplo, sexo: 0 e 1. A gente só precisa vir aqui em cima e ver o que significa o 0 e 1. Aqui. O 0 é masculino, o 1 feminino. Perfeito?

[5:33] Então pessoal, esse primeiro vídeo é isso. No próximo a gente já começa colocando a mão na massa de verdade - a gente vai fazer um teste de hipóteses de primeira.

[5:42] Ok? Só para a gente assustar um pouco. Teste de normalidades, próximo vídeo. Até lá.