**03 - Ambiente e conhecendo o dataset**

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

[0:09] Eu vou utilizar, como fiz nos outros cursos, o Colab. Vamos pesquisar no Google "Colaboratory" e a primeira opção, [disponível neste link](https://colab.research.google.com/?utm_source=scs-index), é que acessaremos.

[0:18] O objetivo é verificar as versões que eu estou utilizando agora, quando eu estou gravando, para quando você estiver fazendo o curso, no futuro, caso apareça algum problema na execução de algum código, você verifica a versão que você está usando. Se for diferente da minha, pode ser esse problema e você pode voltar uma versão ou duas, sem o menor problema.

[0:36] No primeiro curso, vídeo de apresentação, eu mostro como fazer essa volta de versão.

[0:42] Então vou apertar "Esc", 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] No Colaboratory, selecionarei "File" no canto esquerdo superior da página. Em seguida, selecionamos "Upload notebook". Na próxima tela, na aba "Upload" direto, vamos apertar o botão "escolher arquivo". Na próxima tela, abriremos o arquivo "Versão\_bibliotexas.ipynb". Ele vem configurado de maneira que facilita verificarmos as bibliotecas que estamos utilizando.

[1:11] Vou rodar clicando em "Shift + Enter". Ele vai rodar de novo, o meu já tinha rodado. Está executando, demora um pouquinho. Ok. Feito!

[1:20] Estamos usando a versão do Pandas e essas são as bibliotecas que eu vamos utilizar: Pandas 0.24.2; Numpy 1.16.3; Scipy 1.2.1; Statsmodels 0.9.0.

import pandas, numpy, scipy, statsmodels, seaborn, matplotlib

print('Versão do pandas -> %s' % pandas.\_\_version\_\_)

print('Versão do numpy -> %s' % numpy.\_\_version\_\_)

print('Versão do scipy -> %s' % scipy.\_\_version\_\_)

print('Versão do statsmodels -> %s' % statsmodels.\_\_version\_\_)

print('Versão do seaborn -> %s' % seaborn.\_\_version\_\_)

print('Versão do matplotlib -> %s' % matplotlib.\_\_version\_\_)

import pandas, numpy, scipy, statsmodels, seaborn, matplotlib

print('Versão do pandas -> %s' % pandas.\_\_version\_\_)

print('Versão do numpy -> %s' % numpy.\_\_version\_\_)

print('Versão do scipy -> %s' % scipy.\_\_version\_\_)

print('Versão do statsmodels -> %s' % statsmodels.\_\_version\_\_)

print('Versão do seaborn -> %s' % seaborn.\_\_version\_\_)

print('Versão do matplotlib -> %s' % matplotlib.\_\_version\_\_)

COPIAR CÓDIGO

[1:37] Temos, mais ao final, algumas bibliotecas de visualização de gráficos, estudaremos uma coisinha ou outra, é pouca coisa. O Seaborn 0.9.0 e o Matplotlib 3.0.3.

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

[1:55] Eu já deixei aberto o Notebook e nós vamos começar a trabalhar. Ele já está todo organizado e nós vamos começar a discutir sobre teste de hipótese.

[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 com os exercícios, utilizar esse Dataset para ficar um projeto bem legal.

[2:20] Abrindo o tópico "Conhecendo os dados" que está no Colab, o Dataset é o mesmo que a gente temos utilizado 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 a elaborei a partir de uma variável aleatória normal para fins didáticos, isto é, para entendermos o funcionamento de algumas coisas.

[2:49] Temos também "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 procedimento simples.

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

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

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

import pandas as pdCOPIAR CÓDIGO

[3:43] Para economizar, vamos utilizar o Numpy em alguns pontos do nosso treinamento, então eu já vou importar o Numpy também, import numpy as np. Estamos usando esses dois apelidos que são comuns na comunidade: "pd" para Pandas e "np" para Numpy. Ok?

import pandas as pd

import numpy as np

COPIAR CÓDIGO

[4:05] Agora basta apertar "Shift + Enter". Rodou, Ok. Vamos criar um Dataframe com os dados que eu esqueci de mostrar. Abrindo aba lateral do nosso notebook, selecionando Files - nos outros cursos a gente fez isso e o meu já está aberto aqui, por isso que eu esqueci - eu posso acessar Upload. Os dados já estão disponíveis para você fazer download. Seguindo, na próxima tela, selecionaremos o arquivo "dados.csv" e agora poderemos utilizá-lo.

[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. Ok. Vamos criar uma variável dados e vou chamar o Pandas, pd e a função read.csv(), para ler o arquivo CSV e colocá-lo dentro de um Dataframe, e vou passar aqui dados.csv, que é o arquivo que fizemos o Upload.

dados = pd.read\_csv('dados.csv')

COPIAR CÓDIGO

[5:01] Tudo certo. Vamos apertar "Shift + Enter". Rodou. Vamos visualizar os cinco primeiros registro com dados.head() e apertando "Shift + Enter". Está aqui, o Dataset já está carregado com UF, sexo, idade, cor, estudo, renda e altura.

|  | **UF** | **Sexo** | **Idade** | **Cor** | **Anos de Estudo** | **Renda** | **Altura** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 11 | 0 | 23 | 8 | 12 | 800 | 1.603808 |
| 1 | 11 | 1 | 23 | 2 | 12 | 1150 | 1.739790 |
| 2 | 11 | 1 | 35 | 8 | 15 | 880 | 1.760444 |
| 3 | 11 | 0 | 46 | 2 | 6 | 3500 | 1.783154 |
| 4 | 11 | 1 | 47 | 8 | 9 | 150 | 1.1690631 |

[5:19] Alguns dados 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. Vamos fazer um teste de hipóteses de primeira.

[5:42] Ok? Só para a gente assustar um pouco. Teste de normalidades no próximo vídeo. Até lá.