Data: \_\_\_/\_\_\_/\_\_\_\_\_

Horário: 18h50min às 22h00min

Peso: 0,0

Aluno (a): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1) Para o código abaixo faça um breve resumo explanando a sua compreensão lógica de cada linha do código. Utilize da numeração definida em cada linha caso precise reforçar pontos específicos.

**Linha 01:** import pandas as pd

**Linha 02:** import random

**Linha 03:** import numpy as np

**Linha 04:** dataset = pd.read\_csv('census.csv')

**Linha 05:** dataset.shape

**Linha 06:** dataset.head()

**Linha 07:** dataset.tail()

**Linha 08:** df\_amostra\_aleatoria\_simples = dataset.sample(n = 100, random\_state = 1)

**Linha 09:** df\_amostra\_aleatoria\_simples.shape

**Linha 10:** df\_amostra\_aleatoria\_simples.head()

As 3 primeiras linhas são apenas para importar as bibliotecas que seram usadas, linha 4 está lendo um csv e atribuindo a uma variável, linha 5 mostra o formato do dataset, linha 6 as 5 primeiras linhas do dataset, linha 7 mostra as 5 ultimas linhas do mesmo, linha 8 está pegando do dataset(csv) pegando 100 itens aleatórios com o Random\_state igual a 1 para ser possível reproduzir a mesma aletóriedade depois e nas linhas seguintes faz a mesma coisa da linha 5 mostrando o formato e depois pegando as 5 primeiras linhas.

Para o código abaixo faça um breve resumo explanando a sua compreensão lógica de cada linha do código. Utilize da numeração definida em cada linha caso precise reforçar pontos específicos.

**Linha 01:** import pandas as pd

**Linha 02:** import random

**Linha 03:** import numpy as np

**Linha 04:** dataset = pd.read\_csv('census.csv')

**Linha 05:** len(dataset) // 100

**Linha 06:** random.seed(1)

**Linha 07:** random.randint(0, 325)

**Linha 08:** np.arange(68, len(dataset), step = 325)

As 3 primeiras linhas são apenas para importar as bibliotecas que seram usadas, na linha está sendo lido o arquivo csv e atribuído a uma variável, linha 5 pegando quantas linhas tem no dataset e dividindo por 100, linha 6 pegando uma seed randomizada com o código de randomização 1, podendo ser usado para ter o mesmo resultado no futuro, na linha 7 pega um número inteiro entre 0 a 325 e na linha 8 começando da linha 68 indo até o final do dataset pulando de 325 itens de cada item

Para o código abaixo faça um breve resumo explanando a sua compreensão lógica de cada linha do código. Utilize da numeração definida em cada linha caso precise reforçar pontos específicos.

**Linha 01:** import pandas as pd

**Linha 02:** import random

**Linha 03:** import numpy as np

**Linha 04:** dataset = pd.read\_csv('census.csv')

**Linha 05:** len(dataset) / 10

**Linha 06:** grupos = []

**Linha 07:** id\_grupo = 0

**Linha 08:** contagem = 0

**Linha 09:**for \_ in dataset.iterrows():

**Linha 10:**  grupos.append(id\_grupo)

**Linha 11:**  contagem += 1

**Linha 12:**  if contagem > 3256:

**Linha 13:**  contagem = 0

**Linha 14:**  id\_grupo += 1

**Linha 15:** print(grupos)

**Linha 16:** np.unique(grupos, return\_counts=True)

**Linha 17:** np.shape(grupos), dataset.shape

**Linha 18:** dataset['grupo'] = grupos

**Linha 19:** dataset.head()

**Linha 20:** dataset.tail()

**Linha 21:** random.randint(0, 9)

**Linha 22:** df\_agrupamento = dataset[dataset['grupo'] == 7]

**Linha 23:** df\_agrupamento.shape

**Linha 24:** df\_agrupamento['grupo'].value\_counts()

As 3 primeiras linhas são apenas para importar as bibliotecas que seram usadas, na linha está sendo lido o arquivo csv e atribuído a uma variável, linha 5 pega o tamanho do dataset e divide por 10, linha 6 atribui uma lista vazia a uma variável, linha 7 seta uma variável = 0, mesma coisa na 8, na linha 9 cria um for loop para percorrer por cada item do dataset, na linha 10 adiciona um id\_grupo na lista grupos, linha 11 adiciona e atribui a var contagem +1, linha 12 verifica se a contagem é maior 3256 para percorrer o próximo trecho de código, linha 13 seta 0 na contagem, linha 14 adiciona + 1 e coloca em id\_grupo, linha 15 mostra no console o valor de grupos, linha 16 pega cada item único e ve quantas vezes ele aparece na lista, linha 17 mostra quantos elementos em grupos e mostra quantas linhas e colunas no dataset, linha 18 cria um dataset com a lista de grupo, linha 19 mostra as 5 primeiras linhas do dataset, linha 20 as 5 ultimas linhas do dataset, linha 21 pega uma valor aleatório entre 0 a 9, linha 22 pega o dataset apenas onde a coluna grupo é igual a 7 e atribui a uma variável, linha 23 mostra o tipo do dataset e a linha 24 seleciona a coluna e conta quantas vezes aparece no dataset

Para o código abaixo faça um breve resumo explanando a sua compreensão lógica de cada linha do código. Utilize da numeração definida em cada linha caso precise reforçar pontos específicos.

**Linha 01:** import pandas as pd

**Linha 02:** import random

**Linha 03:** import numpy as np

**Linha 04:** dataset = pd.read\_csv('census.csv')

**Linha 05:** def amostragem1(dataset, amostras):

**Linha 06:** intervalo = len(dataset) // amostras

**Linha 07:** random.seed(1)

**Linha 08:** inicio = random.randint(0, intervalo)

**Linha 09:** indices = np.arange(inicio, len(dataset), step = intervalo)

**Linha 10:** amostraaa = dataset.iloc[indices]

**Linha 11:** return amostraaa

Linha 01, 02 e 03 importa as bibliotecas para serem usadas no código, na linha 04 pega o dataset do csv e atribui a variável, na linha 5 monta uma funçã com o nome de amostragem1 tendo as propriedades para serem passadas como dataset e amostras, na linha 6 atribui a uma var o tamanho do dataset e divide por amostras, linha 7 pega uma seed randomizada com o valor 1 para ser passado depois, linha 8 seta uma var com um inteiro randomizado entre 0 ao valor do intervalo, linha 9 seta uma var com os valores pegando as colunas do dataset iniciando pela var do inicio até o tamanho do dataset a cada valor do intervalo, linha 10 seta os valores das linhas em relação aos indeces que estão dentro da var índices e na linha 11 retorna essa var

Para o código abaixo faça um breve resumo explanando a sua compreensão lógica de cada linha do código. Utilize da numeração definida em cada linha caso precise reforçar pontos específicos.

**Linha 01:** import pandas as pd

**Linha 02:** import random

**Linha 03:** import numpy as np

**Linha 04:** dataset = pd.read\_csv('census.csv')

**Linha 05:**from sklearn.model\_selection import StratifiedShuffleSplit

**Linha 06:** dataset['income'].value\_counts()

**Linha 07:** 7841 / len(dataset), 24720 / len(dataset)

**Linha 08:** 100 / len(dataset)

**Linha 09:** split = StratifiedShuffleSplit(test\_size=0.0030711587481956942) **Linha 10:** for x, y in split.split(dataset, dataset['income']):

**Linha 11:**  df\_x = dataset.iloc[x]

**Linha 12:**  df\_y = dataset.iloc[y]

**Linha 13:** df\_x.shape, df\_y.shape

**Linha 14:** df\_y.head()

**Linha 15:** df\_y['income'].value\_counts()

Primeiras 3 linhas pega as bibliotecas que seram usadas no código, linha 4 seta uma var com o conteúdo do csv, linha 5 pega mais um import para ser usado no código, linha 6 pega a linha income do csv e conta quantas vezes aparecem cada valor nela, linha 7 pega o valor 7841 e divide pelo tamanho do dataset e depois faz o mesmo com o valor 24720, linha 8 pega o valor 100 e divide pelo tamanho do dataset, linha 9 seta o valor var para ser usada em uma pesquisa estratificada, linha 10 faz um for para depois fazer a pesquisa estratificada, linha 11 e 12 seta os valores x e y com valores do dataset, pelos índices do for x e y, linha 13 mostra os dois valores setados, linha 14 pega os 5 primeiros itens e na linha 15 pega a lista y com a coluna income conta todos os valores diferentes

Exercício de Estatística Aplicada: Técnicas de Amostragem com o DataFrame da Netflix. Utilize o DataFrame carregado com os dados da Netflix. Nosso objetivo é de aplicar os métodos de amostragem aleatória simples e estratificada para extrair subconjuntos representativos dos dados.

Dica: para abrir este arquivo com sucesso utilize a seguinte codificação:

**df = pd.read\_csv('netflix.csv', encoding='latin1')**

* Em seguida verifique a quantidade de linhas e colunas deste arquivo:
* Valide também quais são os títulos das colunas:

Após realizar os passos anteriores pudemos verificar que o arquivo está um pouco poluído com dados sem valor, certo?

Portanto:

* 1. Limpeza inicial dos dados: elimine todas as colunas Unnamed, pois estão vazias ou são irrelevantes.
  2. Remova linhas com valores nulos nas colunas type, country ou rating.
  3. Selecione 50 registros aleatórios do DataFrame total – aplicando, portanto, amostragem aleatória simples.
  4. Aplicar uma amostragem estratificada na coluna type, selecionando uma amostra estratificada que represente 5% do DataFrame.