CENTRO UNIVERSITÁRIO SATC
ENGENHARIA DE SOFTWARE/COMPUTAÇÃO - DISCIPLINA DE ESTATÍSTICA APLICADA
Professor: Max Gabriel Steiner
Avaliação: Técnicas de Amostragem Aplicadas em Python

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()

2) 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)

3) 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()

4) 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

5) 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()

6) 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.