Centre for infrastructure, Sustainable Transportation and Urban Planning

Indian Institute of Science (IISc), Bengaluru
Summer Internship Program 2024

Follow the instructions below precisely.

- Plagiarism will result in instant disqualification. You must write your own code.
- To make your submission, use the following Google form: https://forms.gle/88bPsuHJSPaCnHdY7
- You are allowed to make only one submission for this test. While submitting, you will be asked to upload three documents:
  - A Python file (format: .py) containing all the codes related to the question.
  - An output file (format: .txt) from the above code
  - A Pseudo code (format: .doc): A pseudo code highlighting the approach.
- If possible, the Python code should be written according to PEP 8 – style guide (reference).
- Your submissions will be evaluated based on your approach and the python codes.
- The test commences on 22nd March 2024 (10:00 AM). The last date for submission is 24th March 2024 (05:00 PM). Late submissions will not be accepted.
- If selected, it will be mandatory for you to join in an in-person capacity. Please refrain from attempting the test if you cannot attend in-person.
- For any clarifications, contact Anil Koushik (anil.koushik@fsid-iisc.in)

All the best.
CiSTUP
IISc, Bengaluru

Question

Consider the dataset provided in the file "Data.csv" (https://drive.google.com/file/d/11GedB2pRmIpjMihPQfQhtroezpahFRB4/view?usp=sharing) which contains a sample of 200 individuals, each characterized by three attributes: sex, age group, and highest education level. The categories for each variable are outlined in Table 1.

Table 1. Description of categories in variables

Variable	Category	Description
Sex	1	Male
	2	Female
Age_group	1	Below 22 years
	2	22-60 years
	3	Above 60 years
Highest_education_level	0	No formal education
	1	Primary education
	2	Secondary education
	3	Graduation and above

The above sample is representative of a region whose population characteristics for the three attributes are provided in Table 2. The table provides total number of individuals (frequencies) in different categories of each of the three variables.

Table 2: Population characteristics

Variable	Description	Frequency
Sex	Male	25324
	Female	24676
Age_group	Below 22 years	17955
	22-60 years	29642
	Above 60 years	2403
Highest_education_level	No_formal education	7490
	Primary education	5655
	Secondary education	24400
	Graduation and above	12455

Using the above seed sample (from Data.csv) and synthesize a population of 50,000 agents such that the generated population matches the actual population characteristics (from Table 2). You need to provide the logic (in the form of a pseudo code) and its python implementation. Specifically:

i. Compose a concise approach for tackling the problem. Write a pseudocode highlighting the approach into a *.doc* file.
ii. Develop a Python code to synthesize the population using the provided sample. At the conclusion of the code, include a segment to compute frequencies for the specified variables, adhering to the format outlined in Table 2. Ensure comprehensive commenting throughout the code to elucidate the logic employed at each step. The entire code must be written in a single *.py* file
iii. Save the outputs of the code (frequencies for the synthesized population) into a *.txt* file.