Assignment 6: Data Privacy
15-316 Software Foundations of Security and Privacy

1. **Private statistics (25 points).**

   Suppose that you are tasked with releasing statistics about a dataset with the following schema.

Feature	*age*	*gender*	*marital_status*	*college_education*	*salary*
Encoding	$\texttt{int} \in [0, 100]$	$\texttt{int} \in \{0, 1\}$	$\texttt{int} \in \{0, 1\}$	$\texttt{int} \in \{0, 1\}$	$\texttt{int} \in [0, 10^6]$

   Note that the particular values used to represent each attribute are not important for this problem, but the fact that each attribute has an upper and lower bound may be. If you would like to assume a particular encoding when explaining your answer (e.g., `gender` $= 0$ for "male"), then please state your assumed encoding.

   Part 1 (5 points). Show that if you are allowed to query the dataset by counting the number of entries with a particular set of values, then it is possible to learn a person's salary. In particular, you have access to the following function, which you can query as many times as you like:

   $$\texttt{count}(\mathit{age}, \mathit{gender}, \mathit{education}, \mathit{marital}, \mathit{salary}) = |\{\#\text{ database rows matching given values}\}|$$

   Your aim is to learn the salary of a particular individual for whom you know all attributes *except* salary, and you may assume knowledge of the rest of the dataset as described in lecture for the differential privacy threat model.

Part 2 (10 points). Now your goal is to provide statistics about the average salary across gender and education level while satisfying ϵ-differential privacy.

- You have access to the database through variable X, which you should assume is an array containing N dictionaries that you can index by attribute name; i.e., `X[0]["salary"]` returns the salary of the first row of the database.
- You may call a function `Laplace(b)`, which returns a single random sample from the zero-centered Laplace distribution with scale parameter b:

$$\Pr[\texttt{Laplace}(b) = x] \propto \frac{1}{2b} e^{-|x|/b}$$

- You may assume that the breakdown of X by gender and education level is not private information.

Explain how to implement a 1-differentially private function `mean_by_gender_and_edu`, which returns a 4-tuple of floats containing the mean salary for each gender and education level in X. That is, this function privately computes the following statistics:

$$\texttt{mean\_by\_gender\_and\_edu} = (\text{mean}(\textit{women}), \text{mean}(\textit{men}), \text{mean}(\textit{college}), \text{mean}(\textit{nocollege}))$$

Be sure to state which composition principles your solution uses, if any. If it is easiest to present your solution as pseudocode then please do so, but you should explain how it works in words as well.

Part 3 (10 points). In order to support an *arbitrary* number of future queries on this data while still achieving privacy, you decide to make a differentially private histogram to represent it. In other words, you process the data in the following way.

1. Discretize age by bucketing it into 25 year increments, so now each row has an *age* feature taking a value in $\{0, 1, 2, 3\}$.
2. Discretize salary by bucketing it into increments of \$25,000, so now each row has a *salary* feature taking a value in $\{0, 1, 2, 3\}$.
3. Construct a 5-dimensional array H by counting the database as follows:

$$H[i, j, k, l, m] = |\{x \ : \ x[age] = i, x[gender] = j, x[marital] = k, x[college] = l, x[salary] = m\}|$$

Explain how to publish H in a manner that satisfies 1-differential privacy. The private version of H should look like a normal histogram: a table of natural numbers. Then, explain how you can use the private histogram to approximate the `count` function from Part 1. *Note: there are multiple ways to answer the last part of this question, and you will receive credit for a sub-optimal approach as long as you identify the factors that might introduce approximation error.*