Unity ID: asinha6

Question 1:

a) As has been taught in class, in the context of differential privacy, the global sensitivity of a function indicates how much the function's output can change when a single entry in the dataset is added or removed.

Assuming that the original dataset is D, and the adjacent dataset is D'. Also assuming that the number of values in the dataset is 'n'.

Global Sensitivity can be defined as the maximum amount by which its output can change when we add or delete a single element from the dataset.

It can be mathematically represented as:

$$GS_f = max(f(x) - f(x'))$$

In the question, f(x) is the mean which returns the average salary in the dataset. Similarly, f(x') is the mean of the adjacent dataset.

In this scenario, we want to identify the largest change in average salary when one employee's salary is replaced by another, that is, when one salary (x) is removed and another (y) is added while keeping the dataset size N constant.

Summarizing:

1. Finding the mean f(x) for dataset D
2. Removing one random value from dataset D to get dataset D'
3. Calculating f(x')
4. Finally, finding the maximum of f(x) - f(x')

Step 1:

f(x) can be calculated as follows: $\frac{\sum_{i=1}^{n} x_i}{n}$

Step 2:

Now we must pick which value should be deleted in order to obtain the maximum difference between the two datasets. We know that the dataset's value range is [a, b]. We must first check by eliminating the maximum value from the dataset, followed by deleting the minimum value from the dataset. Because the dataset's range is [a, b], the maximum value is b and the minimum value is a.

Step 3:

In case 1: Remove 'a', the f(x') = $\frac{\sum_{i=1}^{n} x_i - a}{n-1}$

In case 2: Remove 'b', the f(x') = $\frac{\sum_{i=1}^{n} x_i - b}{n-1}$

Step 4:

So Case 1: Remove ‘a’

$$sensitivity_a = abs(\frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} x_i - a}{n-1})$$

Case 2: Remove ‘b’

$$sensitivity_b = abs(\frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} x_i - b}{n-1})$$

Now, how do we select the sensitivity? We can take the maximum of both the above cases: Therefore,

$$sensitivity = max[abs(\frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} x_i - a}{n-1}),\ abs(\frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} x_i - b}{n-1})]$$

b) We may add noise to the output using the Laplace distribution (μ = 0, λ). The value of should be as follows to ensure eps - D:

$$\lambda \geq \frac{|h - h'|}{\epsilon} \geq \frac{abs(\frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} x_i - a}{n-1})}{\epsilon}$$

Or,
$$\geq \frac{abs(\frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} x_i - b}{n-1})}{\epsilon}$$

The above equations have been derived from part a, where sensitivity was calculated. The sensitivity was calculated as:

$$sensitivity = max[abs(\frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} x_i - a}{n-1}),\ abs(\frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} x_i - b}{n-1})]$$

We've already calculated the mean function's global sensitivity. To set the Laplacian distribution parameter for the mean function, we may simply apply the formula:

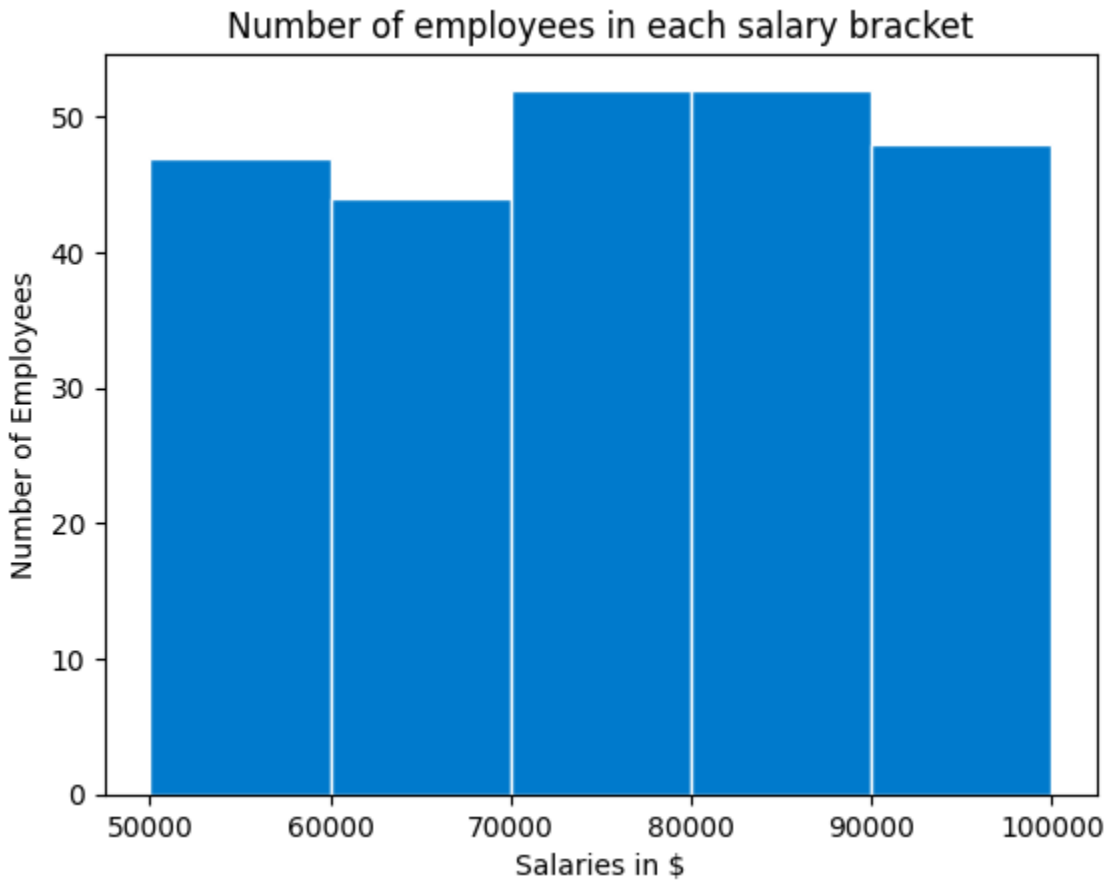
$$\lambda = \frac{sensitivity}{\epsilon}$$

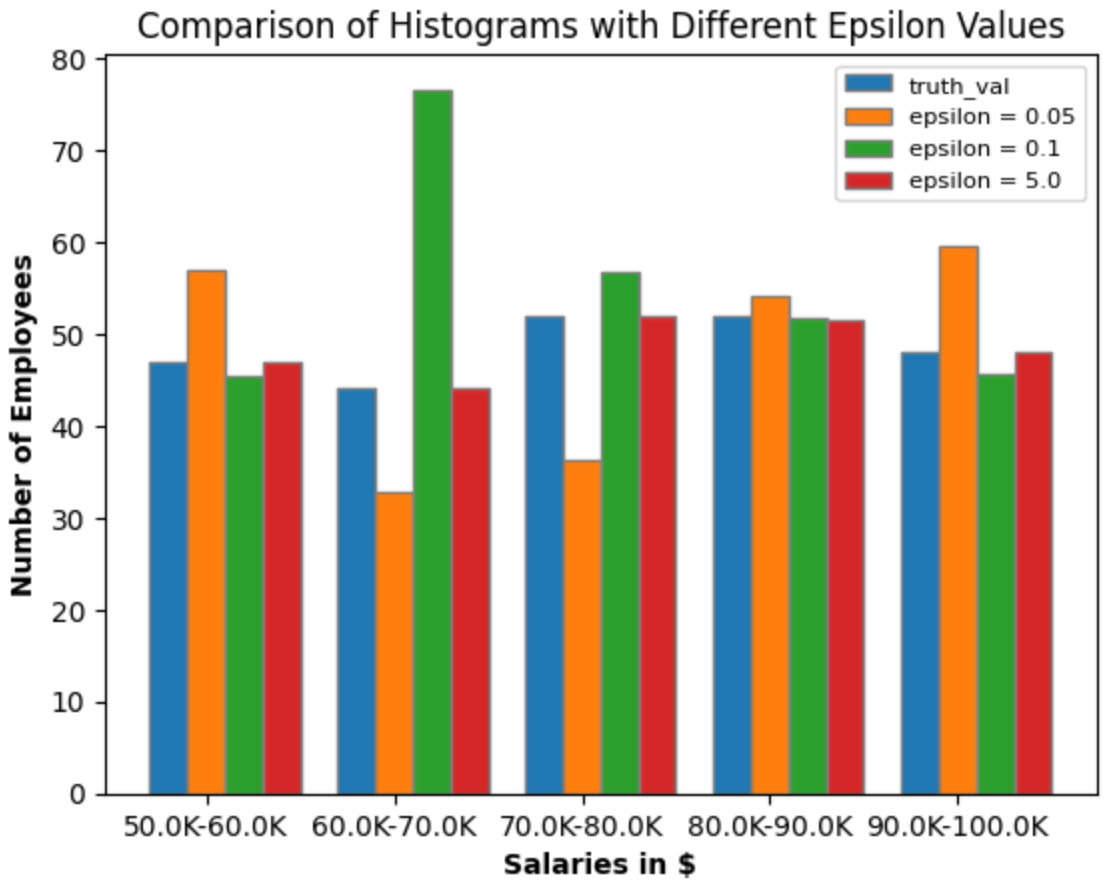
To get the required level of privacy, we must make a decision. A smaller **ε** gives greater privacy protections, but it may introduce more noise into the output, making it less precise. A greater value of ε, on the other hand, gives weaker privacy assurances while adding less noise to the output, resulting in a more accurate but less private calculation.

So, in order to configure the Laplacian distribution parameter λ to provide differential privacy for the mean function, we must first decide on the privacy parameter ε that corresponds to the desired privacy level and acceptable trade-off between privacy and accuracy. Then, using the above technique, we may compute the suitable based on the calculated global sensitivity (f).

Question 2

a)





b) We can see that when we increase the value of ϵ, the degree of the change in noise diminishes based on the random value generated for smaller values of ϵ. When ϵ is 0.05, the noise is high, and when ϵ is between 0.1 and 5.0, the magnitude of the noise decreases.
In other words,
As we raise ε, we find that the output perturbation due to noise, i.e. the distortion in the histogram, reduces and begins to resemble the original dataset. The privacy of the dataset reduces as increases.
As ε increases, so does the utility of the histogram. This is because the distortion lessens as ε increases, and it resembles the original dataset. When = 5.0, the histogram is identical to the original, there is no disturbance, and the utility is maximized.

Question 3

To solve the 16-bit database problem in order to retrieve the i-th (8th) bit using a 3-server PIR (Private Information Retrieval) protocol under O(n1/2) scheme, we need to follow the following steps:

Step 1: Convert the 16-bit data into a 2-D matrix. The matrix looks as follows:

0	1	1	0
0	1	0	1
1	0	0	1
1	0	1	1

Since, the user is interested in i-th (8th bit), in the 2-D matrix, [i,j] = [2,0]

Step 2: For Server S1, we need to select Q1, a random string. Let Q1 = {1010}. Now, we must take the dot product of Q1 with each row. The dot product looks as follows:

0		1	
0		0	
1		0	
1		1	

Step 3: Now, we must perform component bit-wise XOR. The final result looks as follows:

1
0
1
0

The Server S1 will therefore send 1010 as result to the user.

Step 4: For server S2, we need to select another random query string. Let Q2 = {1001}. Let's perform dot product and then perform XOR on the dot product

				XOR Result
0			0	0
0			1	1
1			1	0
1			1	0

The Server S2 will therefore send 0100 as result to the user.

Step 5: To find Q3 for server S3, we must perform XOR operation between Q1, Q2 and {j}
Note: {j} is a 4 bit string of 0s with the jth position's bit flipped. (j=0)
Q3 = Q1⊕Q2⊕{i} = 1010 ⊕ 1001 ⊕ 1000 = {1011}

Step 6: We must now find the dot product and then XOR the result.

				XOR Result
0		1	0	1
0		0	1	1
1		0	1	0
1		1	1	1

The Server S3 will therefore send 1101 as the result to the user.

Step 7: The user will receive result strings from S1, S2, S3. The highlighted element was present in row i ⟹ 2. Thus, the user will extract the bit at i ⟹ 2 from both the results. Therefore, it will extract 1 from S1, 0 from S2 and 0 from S3. Now, the user will xor these values and retrieve the desired result. perform XOR operation on the i-th bit of the result strings that have been sent by S1, S2 and S3.
1010 ⊕ 0100 ⊕1101 = 1⊕0⊕0 = 1.
Therefore, the user finally gets the i-th(8th bit) = 1.
We can conclude that privacy has been maintained because the user successfully obtained the necessary row/column without broadcasting it to the servers.