Privacy Assignment #1

Submitted by : Prashanthi Kanniappan Murthy

1. a. Key Identifiers (KI) : *ID, First Name, Last Name, SSN*

   The above fields are considered Key Identifiers because they uniquely identify a specific person with any one of the columns given.

   b. Quasi-identifiers (QI) : *Address, DoB, City, ZIP, Credit Card Type*

   Any combination of above columns can uniquely identify a specific person.

   c. Sensitive Data (SD) : *Credit Card Number, Expiration Date, CCV, Amount.*

   d. Non-sensitive data(NSD) : This set is empty as table doesn't contain non-sensitive data, exposing any one of the field will have impact in one way or the other.

2. Focussing on DOB and ZIP, achieve 3-anonymity, using generalisation methods. From the table given in the question :

DOB	ZIP
1985	98195
1995	98201
1978	98501
1983	99205
1990	98195
1965	98195

After applying generalisation to the DOB keeping 1985 as a cut off, we'll have 3 anonymity. Table after applying anonymity:

DOB	ZIP	Equivalence Class
≥1985	98***	1
≥1985	98***	1
≥1985	98***	1
<1985	9****	2
<1985	9****	2
<1985	9****	2

The table satisfies 3 anonymity and there are 2 equivalence classes.

3. Focussing on DOB and credit card type, achieve 2-diversity.
Original table :

DOB	Credit Card Type
1985	VISA
1995	VISA
1978	AMEX
1983	MASTER
1990	VISA
1965	VISA

After applying l-diversity , partitioning on 1980 to be the gapping year,

DOB	Credit Card Type
>1980	VISA
>1980	VISA
>1980	VISA
>1980	MASTER
<1980	AMEX
<1980	VISA

4. a. The sensitivity of this function is 1, as a person might be HIV positive or negative. So removing a row from the database can maximum affect by 1 count, since this is a count query.

b. Query Response Algorithm with 0.01 Differential Privacy:

Inputs:

D : Hospital database of patients with a single column indicating whether or not the patient is HIV positive or not.

f : Query ("How many patients are HIV positive?")

ϵ : 0.01

Query Response

$$\gamma = \text{f(D)} + \text{Noise}$$

Replacing Noise,

$$\gamma = \text{f(D)} + \text{Lap}(\frac{\Delta f}{\epsilon})$$

Substituting values for Sensitivity and Epsilon,

$$\gamma = \text{f(D)} + \text{Lap(100)}$$

c. Total Privacy = $\epsilon_{total} = \sum_{i=1}^{q} \epsilon_i$

$$\epsilon_{total} = \sum_{i=1}^{10} \epsilon_i$$

$$\epsilon_{total} = \sum_{i=1}^{10} 0.01$$

$$\epsilon_{total} = 0.01 * 10$$

$$\epsilon_{total} = 0.1$$

5.a. Sensitivity of max function :

To calculate the sensitivity, the entire domain of values possible should be considered. Let x_1, x_2, x_3 be 3 values taken into consideration in the same ascending values.

x_1 - the second largest value present in the database

x_2 - the largest value present in the database

x_3 - the largest value present in the domain, but not in the database

Let *max(*D*)* return x_2 , for D' , let us assume x_2 is removed.

For any value x_i added in the database in the range $[x_1, x_2]$the sensitivity will be $x_2 - x_i$, if just removal, $x_2 - x_1$

For any value x_j added in the range $[x_2, x_3]$ the sensitivity will be $x_j - x_2$, if just removal, $x_3 - x_2$

The Global Sensitivity is *max*($x_2 - x_1$, $x_3 - x_2$)

b. Query Response Algorithm with ϵ - differential privacy:

Inputs:

D : Dataset containing annual salaries of all NCSU employees

$max(D)$ - returns the maximum salary in the dataset.

ϵ

San : Standard Laplacian distribution

Given a *f* , *San* generates random ξ from Laplacian Distribution with variance that depends on the sensitivity of function *f* and the parameter ϵ

Query Response:

γ = f(D) + ξ

γ = max(D) + $Lap(\frac{\Delta f}{\epsilon})$

Taking Sensitivity from the above calculation ,

Δf = *max*($x_2 - x_1$, $x_3 - x_2$)

Substituting back ,

γ = max(D) + $Lap(\frac{max(x_2 - x_1, x_3 - x_2)}{\epsilon})$

6. a. Sensitivity of mean function :

To calculate the sensitivity, the numbers in the database is considered. Let x_i be all the numbers in dataset of *n* values . To get the maximum difference, lets assume the largest value x_l is removed.

Let the sum of the dataset be $S = \sum_{i=1}^{n} x_i$

The sensitivity is thus defined as

$$\Delta f = |\frac{S}{n} - \frac{S - x_l}{n-1}|$$

b. Query Response Algorithm with ϵ - differential privacy:

Inputs:

D : Dataset containing annual salaries of all NCSU employees

$max(D)$ - returns the maximum salary in the dataset.

ϵ

San : Standard Laplacian distribution

Given a *f* , *San* generates random ξ from Laplacian Distribution with variance that depends on the sensitivity of function *f* and the parameter ϵ

Query Response:

$$\gamma = \text{f(D)} + \xi$$

$$\gamma = \text{max(D)} + Lap(\frac{\Delta f}{\epsilon})$$

Taking Sensitivity from the above calculation ,

$$\Delta f = |\frac{S}{n} - \frac{S - x_l}{n-1}|$$

Substituting back,

$$\gamma = \text{max(D)} + Lap(\frac{|\frac{S}{n} - \frac{S - x_l}{n-1}|}{\epsilon})$$

7. Raw data plot :

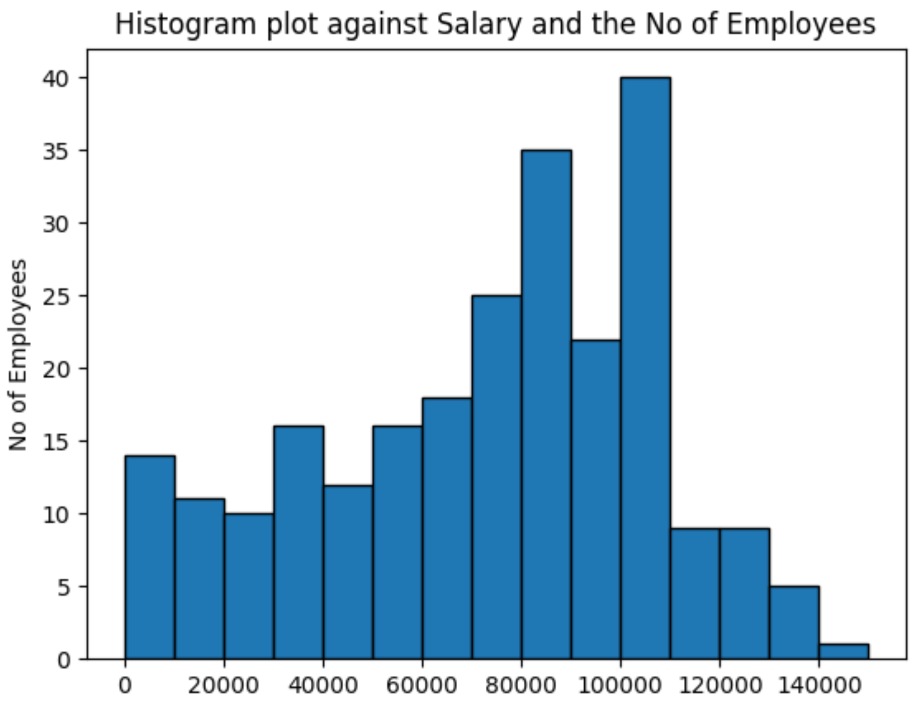

FIG1 : PLOT WITH RAW DATA

Plot with $\epsilon = 0.1$:

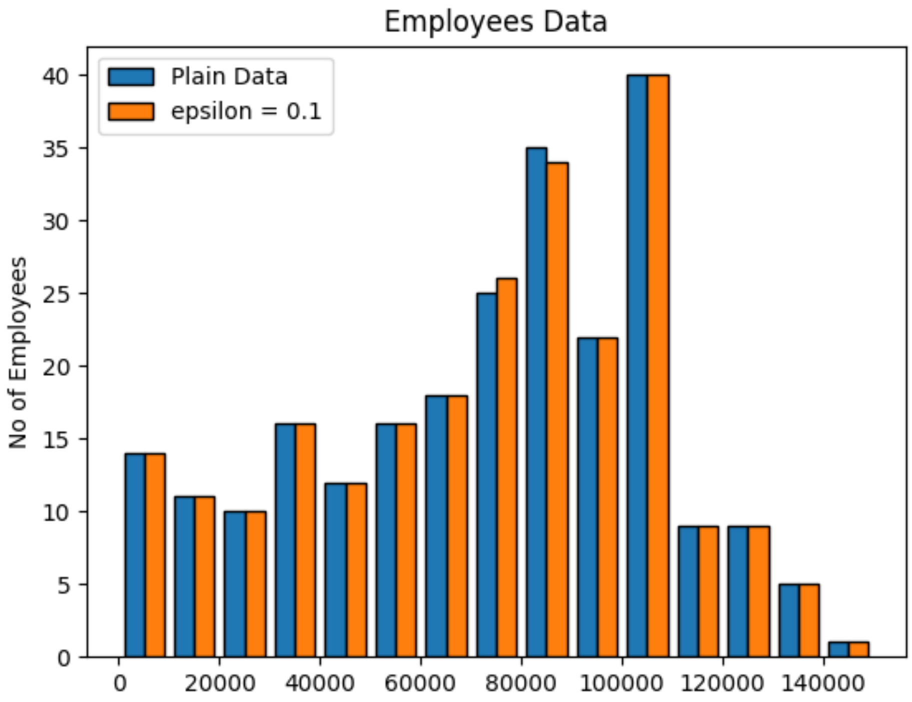

FIG2 : PLOT WITH RAW DATA AGAINST 0.1 DP

Plot with $\epsilon = 1$:

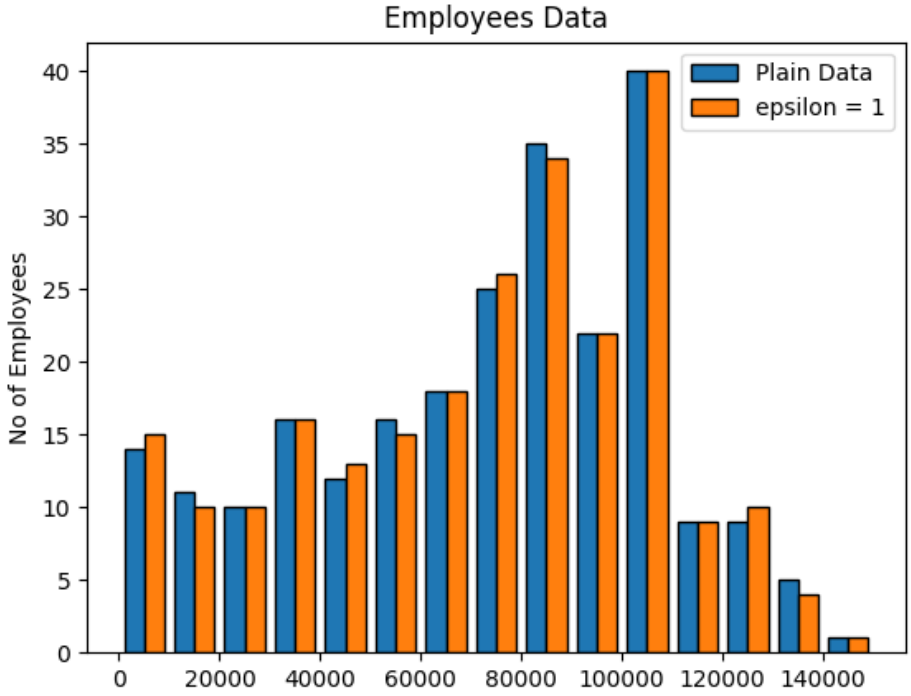

FIG3: PLOT WITH RAW DATA AGAINST 1 DP

Plot with $\epsilon = 10$:

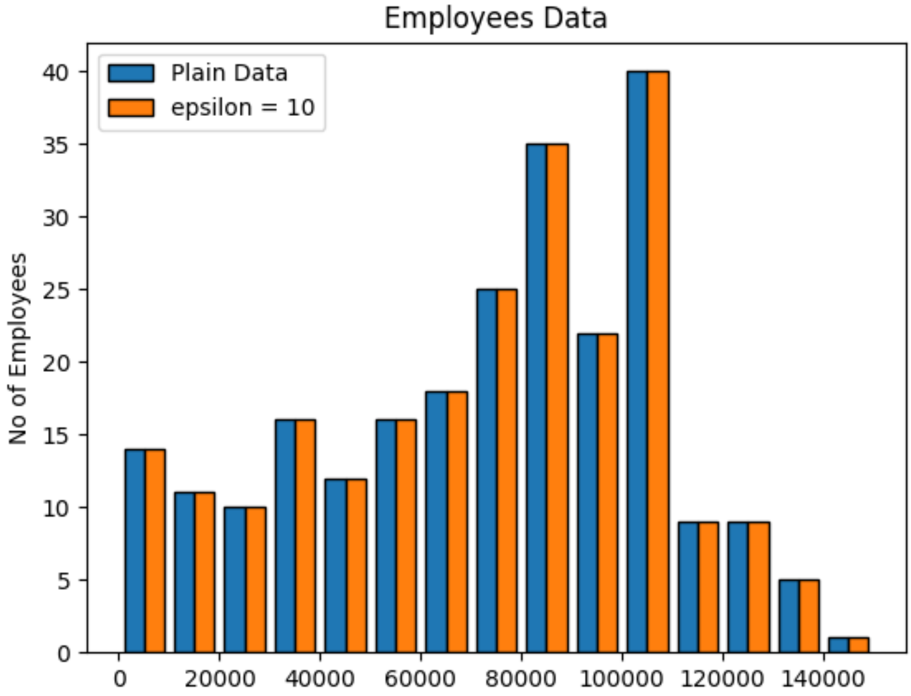

FIG4 : PLOT WITH RAW DATA AGAINST 10 DP

Plot with all the different scales with raw data:

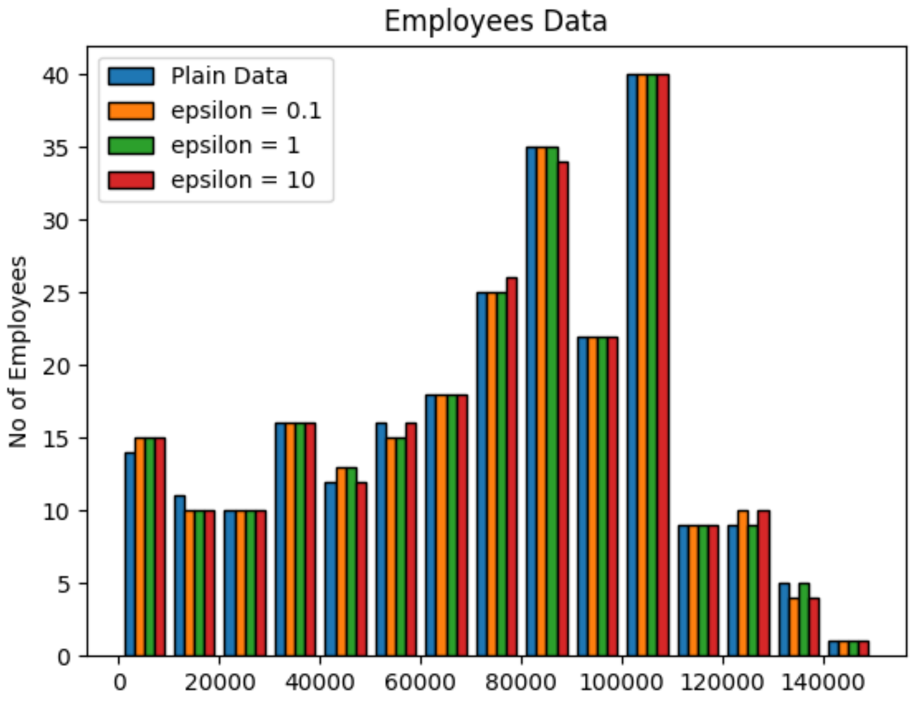

FIG5: PLOST WITH RAW DATA WITH ALL THE DIFFERENT SCALES

As epsilon increases, the distortion of the graph decreases and starts to coincide with the raw data graph. This implies that, as epsilon increases privacy of the dataset decreases.

Utility of the histogram: As epsilon increases, the histogram coincides with the original graph, which implies the utility of the histogram increases. As it can be seen in Fig4, the raw data graph has coincided with the epsilon = 10 data. Hence Utility increases as epsilon increases.