Practical Assignment 02

Security and Privacy

Syntactic and Differential Encryption Models

06/01/2022

Gustavo Valeriano Neves Luizon - uc2021179906

Tiago Conceição – uc2021167993

Summary

Preparation: Dataset and Problem 3
- Firefighters Dataset 3
- Dataset Analysis 3

Syntactic Models 3
- Classifying Attributes 3
- Quasi-Identifiers Evaluation 4
- Privacy Risks of the Dataset - Original Form 5
- Privacy Model and Configuration 6
- Utility and Privacy Results 7
- Dataset Analysis Results 8

Differential Privacy 9

References 9

Preparation: Dataset and Problem

Firefighters Dataset

The U. S. Fire Administration tracks and collects information on the causes of on-duty firefighter fatalities that occur in the United States. The purpose of data collect is to conduct an annual analysis to identify specific problems so that we may direct efforts toward finding solutions that will reduce firefighter fatalities in the future.
The dataset is composed of 2005 fatalities registers and has Name, Age, Rank, Classification, Date of Incident, Cause of Death, Nature of Death, Duty, Activity, Emergency and Property Type information.

Dataset Analysis

The 2 chosen different analysis that are relevant on the dataset are:

1) Firefighters Age Average
2) Firefighters Age Average per Duty

Tabela 1 - Dataset Analysis

-	Dataset Original
Age Average	46,8
Age Avg per Duty - On Duty	50,3
Age Avg per Duty - On-Scene Emergency	51,4
Age Avg per Duty - On-Scene Fire	45,2
Age Avg per Duty - Response	44,9
Age Avg per Duty - Return	50,7
Age Avg per Duty - Training	43,8

Syntactic Models

Classifying Attributes

The “Full Name” attibute is the only one which will Always identify a person in the dataset, so its is classified as “Identifying”, the quasi-identifier attributes can identify persons in dataset individualy or when combined with other quase-identifiers, the quase-identifier attributes are “Age”, “Rank”, “Classification”, “Date of Incident”, “Cause of Death”, “Nature of Death”, “Duty”, “Activity”, “Emergency” and “Property Type”. Finally, the attributes “Cause of Death” and “Nature of Death” are classified as sensitive atributes because they are private attributes that should not be publicly disclosed and may be also linked to identify individuals.

Full Name:
Identifying

Classification:
Quase-Identifying

Date of Incident:
Quase-Identifying

Activity:
Quase-Identifying

Age:
Quase-Identifying

Duty:
Quase-Identifying

Cause of Death:
Quase-Identifying

Emergency:
Insensitive

Rank:
Quase-Identifying

Property Type:
Quase-Identifying

Nature of Death:
Quase-Identifying

Quasi-Identifiers Evaluation

The quase-identifiers can be evaluated by computing the distinction and separation parameters. If we look at the os quasi-identifiers separately, we can see that all attributes have a very high separation value, which indicates that records can be easily differentiated by each attribute. The “Date of Incident” attribute is the only one with a high distinction value, see figure 1.

Quasi-identifier	Distinction	Separation
Emergency	0.11105%	43.65093%
Classification	0.11105%	49.97039%
Duty	0.33315%	71.02924%
Property Type	0.6663%	79.09729%
Activity	1.22154%	87.84422%
Age	4.10883%	94.15134%
Rank	12.15991%	76.97656%
Date of Incident	66.62965%	96.39891%

Figure 1 - QIDs Individualy

Since an attribute looked at separately presents high values of distinct and separation, its combination with other attributes will hardly present better results, therefore, to evaluate the attributes in pairs, the attribute "Date of Incident" was removed. For the evaluation in pairs, it is possible to observe "distinction" values around 30% for the combinations between "Age", "Rank" and "Activity", see figure 2.

Quasi-identifier	Distinction	Separation
Rank, Duty	18.60078%	92.28614%
Rank, Cause Of Death	18.8784%	94.26294%
Age, Cause Of Death	19.37812%	96.00925%
Age, Property Type	20.71072%	96.12789%
Rank, Property Type	20.98834%	94.40971%
Rank, Activity	24.59745%	96.34845%
Age, Activity	34.48084%	96.27004%
Age, Rank	36.59078%	97.75298%

Figure 2 - QIDs in tuples

Finally, it is possible to identify higher distinction values only for the combination of all the remaining attributes after removing "Age", "Rank", "Activity" and "Date of Incident", see figure 3.

Quasi-identifier	Distinction	Separation
Classification, Cause Of Death, Nature Of Death, Duty, Emergency, Property Type	20.71072%	95.51626%
Classification, Cause Of Death, Nature Of Death, Duty, Activity, Emergency	24.20877%	95.77093%
Cause Of Death, Nature Of Death, Duty, Activity, Emergency, Property Type	31.76013%	96.02233%
Classification, Cause Of Death, Nature Of Death, Activity, Emergency, Property Type	32.14881%	96.11265%
Classification, Nature Of Death, Duty, Activity, Emergency, Property Type	33.14825%	96.08958%
Classification, Cause Of Death, Duty, Activity, Emergency, Property Type	33.64797%	96.07675%
Classification, Cause Of Death, Nature Of Death, Duty, Activity, Property Type	36.70183%	96.16392%
Classification, Cause Of Death, Nature Of Death, Duty, Activity, Emergency, Property Type	38.03443%	96.18385%

Figure 3 - QIDs All

Privacy Risks of the Dataset - Original Form

The high distinction and separation values identified indicate the fragility of the dataset in its natural format, the results of the resistance to attack tests confirm the suspicion, presenting high risk rates to the Prosectur, Journalist and Marketer attacks as can be seen in figure 4 and figure 5.

Measure	Value [%]
Lowest prosecutor risk	0.43103%
Records affected by lowest risk	12.88173%
Average prosecutor risk	81.73237%
Highest prosecutor risk	100%
Records affected by highest risk	81.17712%
Estimated prosecutor risk	100%
Estimated journalist risk	100%
Estimated marketer risk	81.73237%
Sample uniques	81.17712%
Population uniques	65.90908%
Population model	ZAYATZ
Quasi-identifiers	Activity, Age, Classification, Date of Incident, Duty, Property Type, Rank

Figure 4 – Dataset Risk for Original Dataset

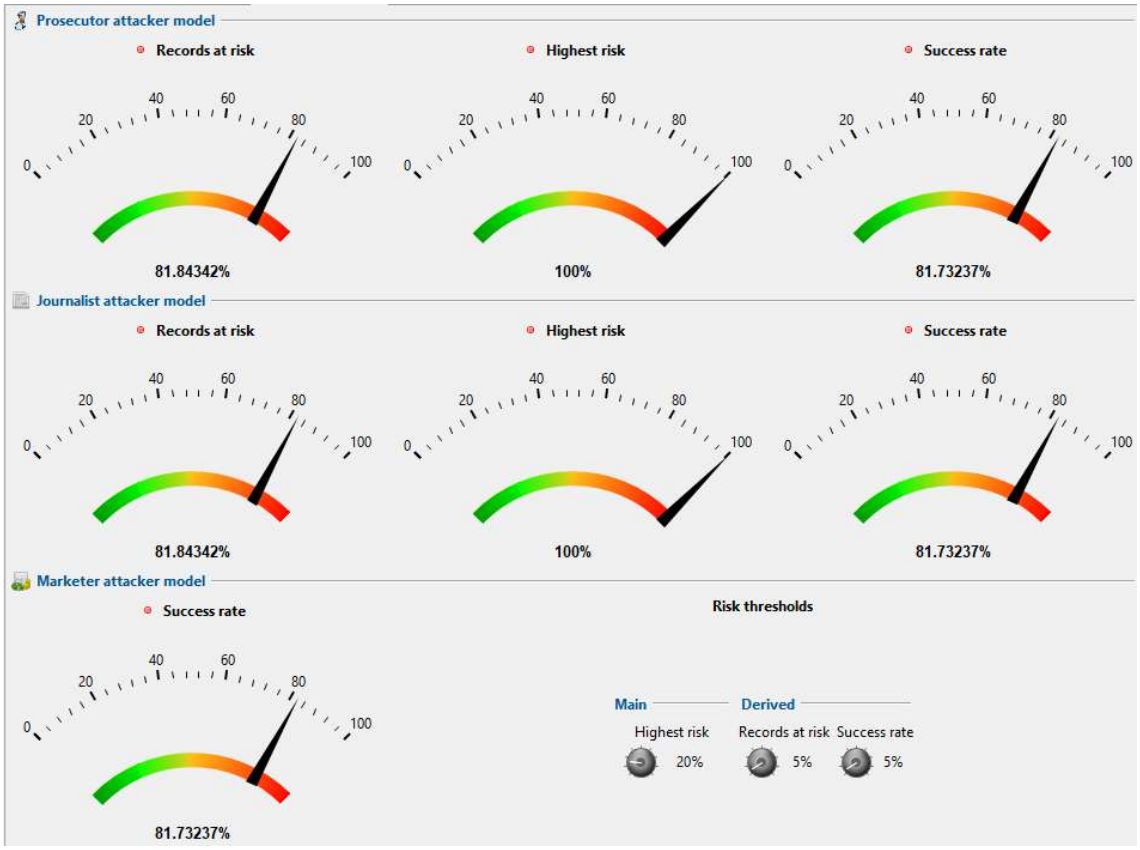

Figure 5 – Risk Thermometers for Original Dataset

Privacy Model and Configuration

To improve the privacy preserving and allow the dataset publishing, the k anonymity Syntactic Privacy Model was applied, with k value equals 5 and 2 different and 2 different grouping configurations in order to obtain the calculations described in item “Dataset Analysis”. The “Cause of Death” and “Nature of Death” attributes were classified as sensitive because individual-specific private attributes that should not be publicly disclosed, may be also linked to identify individuals. Futhermore, a hierarchical generalization structure was defined for each quasi-identifier attribute, as described below:

Age: The age attribute has 5 levels, level 0 is the original value, level 1 makes groups of 10 in 10 years, level 2 makes groups of 20 in 20 years, level 3 of 40 in 40, level 4 divides the records between the group 0 to 60 years old and the group older than 60 years old, finally level 5 deletes all age records.

Rank: The Rank attribute has a very disperse distribution, with a large amount of positions occupied by few people, the grouping hierarchy was carried out in 4 levels, level 0 has the original information, level 1 makes groups of 5 positions, the level 2 does groups by 20, level 3 divides records into 2 groups and level 4 suppresses all records.

Classification: For the classification attribute, the vast majority of records are concentrated in the career and volunteer classifications, so it was more effective to remove records with values different from those of the dataset, since this procedure results in a smaller amount of information loss.

Date of Incident: The attribute "Date of Incident" received grouping levels by month, year and decade, in addition the low amount of records for years prior to 2000 was extremely low, enabling easy recognition of people, these records were removed for showing the procedure which has less loss of information.

Duty: The attribute "Duty" has 2 very representative categories, which are "On-Duty" and "On-Scene-Fire", the others have few records. The level 1 of the hierarchical generalization structure is formed by the 2 main categories and a third grouping containing the other categories, level 2 of the structure is the suppression of records.

Activity: The “Activity” attribute has a very disperse distribution just like the “Rank” attribute, with a large amount of positions occupied by few people, the grouping hierarchy was carried out in 4 levels, level 0 has the original information, level 1 makes groups of 3 positions, the level 2 does groups by 6, level 3 divides records into 2 groups and level 4 suppresses all records.

Property Type: The attribute "Property Type" has 4 very representative categories, which are "Residencial", “Street/Road”, “Outdoor Property” and "Store/Office", the others have few records. The level 1 of the hierarchical generalization structure is formed by the 4 main categories and a fifth group containing the other categories, level 2 of the structure is the suppression of records.

Utility and Privacy Results

The k anonymity Syntactic Privacy Model proved to be efficient in protecting the privacy of people present in the dataset, the vulnerability to Prosecutor, Journalist and Marketer attacks were reduced from values ranging from 80 to 100% in the original dataset to values below 20% as can be seen in figure 6 and figure 7.

Measure	Value [%]
Lowest prosecutor risk	0.20747%
Records affected by lowest risk	26.76291%
Average prosecutor risk	1.27707%
Highest prosecutor risk	16.66667%
Records affected by highest risk	0.33315%
Estimated prosecutor risk	16.66667%
Estimated journalist risk	16.66667%
Estimated marketer risk	1.27707%
Sample uniques	0%
Population uniques	0%
Population model	DANKAR
Quasi-identifiers	Activity, Age, Classification, Date of Incident, Duty, Property Type, Rank

Figure 6 - Dataset Risk for K-Anonymity

Figure 7 - Risk Thermometers for K-Anonymity Model

The data transformation was performed using two different configurations. The first configuration used prioritizes the maintenance of the "Age" attribute because it is an essential attribute for obtaining the calculations described in , the configuration that has the least loss of information considering this assumption is the configuration “1-4-0-3-2-4-2”.

The second configuration used prioritizes the maintenance of the "Rank" attribute, keeping the "Age" attribute with the least change possible, the configuration that has the least loss of information considering this assumption is the configuration “3-4-1-3-1-4-2”, see figure 8.

QID	Config 1	Config 2
Age	1	3
Rank	4	4
Classification	0	1
Date of Incident	3	3
Duty	2	1
Activity	4	4
Property Type	2	2

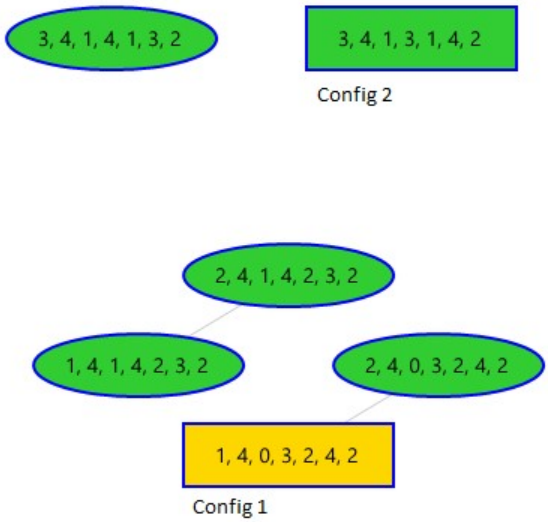

Figura 8 – Transformation Configurations

In order to evaluate the utility of the datasets after the application of the transformations, the averages of the ages of firefighters and the average of the ages of firefighters by occupation at the time of the incident were calculated. The result can be seen in table 2.

Table 2 – Dataset Analysis

-	Original	Config_1	Config_2
Age Average	46,8	45,6	44,5
Age Avg per Duty - On Duty	50,3	45,7	47,5
Age Avg per Duty - On-Scene Emergency	51,4	45,7	41,9
Age Avg per Duty - On-Scene Fire	45,2	45,7	44,5
Age Avg per Duty – Response	44,9	45,7	41,9
Age Avg per Duty – Return	50,7	45,7	41,9
Age Avg per Duty – Training	43,8	45,7	41,9

The results do not show large variations in the mean values, however, we noticed a great loss of information with the groupings, this is because the attributes have a very spread distribution, they have many categories and few records in each category, this causes high level transformations. We see that in configuration 1, the Rank attribute was completely suppressed so that we were able to maintain age group at level 1, and in configuration 2, the level 1 transformation of the "Rank" attribute can only be maintained by raising the transformation of the "Age" attribute to level 3.

Differential Privacy

	0.01	0.2	ln(2)	ln(3)
Mean Age	0.006199	0.016362	0.013865	0.022634

Figure 9 - Differential Privacy applied to the mean age

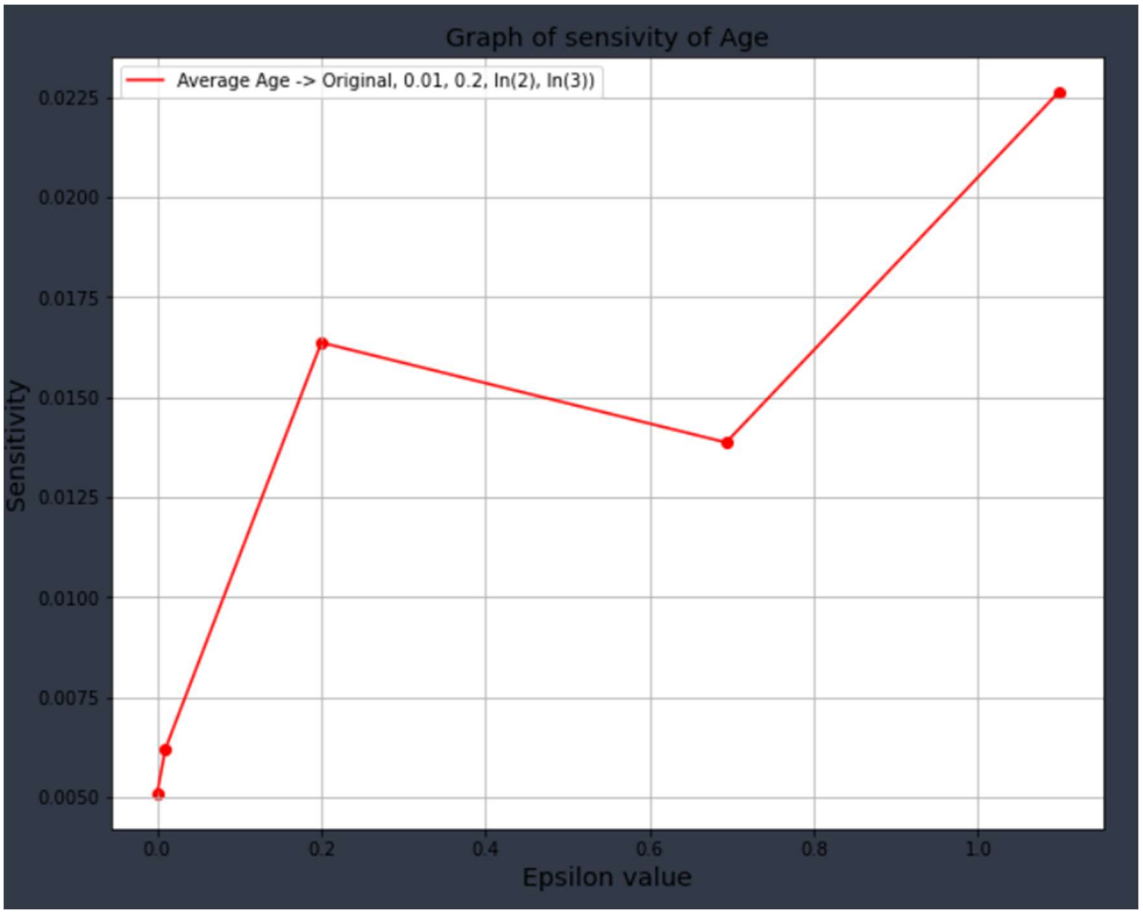

Figure 10 - Graph of the mean age applying differential privacy

	0.01	0.2	ln(2)	ln(3)
Mean age of duty: Response	34.616061	2.891138	3.433103	2.824433
Mean age of duty: On-Scene Fire	6.100689	5.654985	4.928982	5.133583
Mean age of duty: On-Duty	20.229801	7.559591	7.572184	7.989513
Mean age of duty: Training	35.434215	8.446723	8.269744	7.868554
Mean age of duty: On-Scene Emergency	5.716853	10.319073	11.317878	11.522232
Mean age of duty: Return	52.357668	4.066497	6.625205	3.512860

Figure 11- Differential Privacy applied to the mean age of each duty

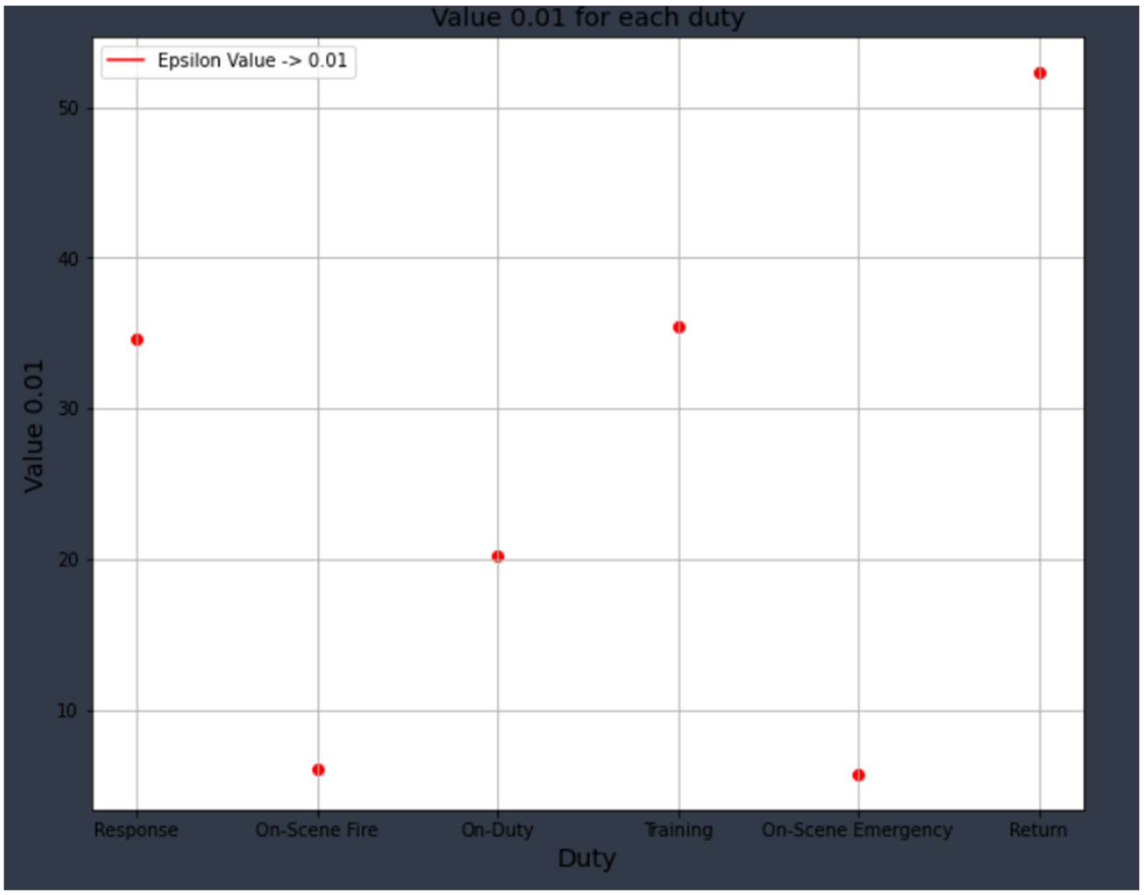

Figure 12 - Differential Privacy for each duty with epsilon value 0.01

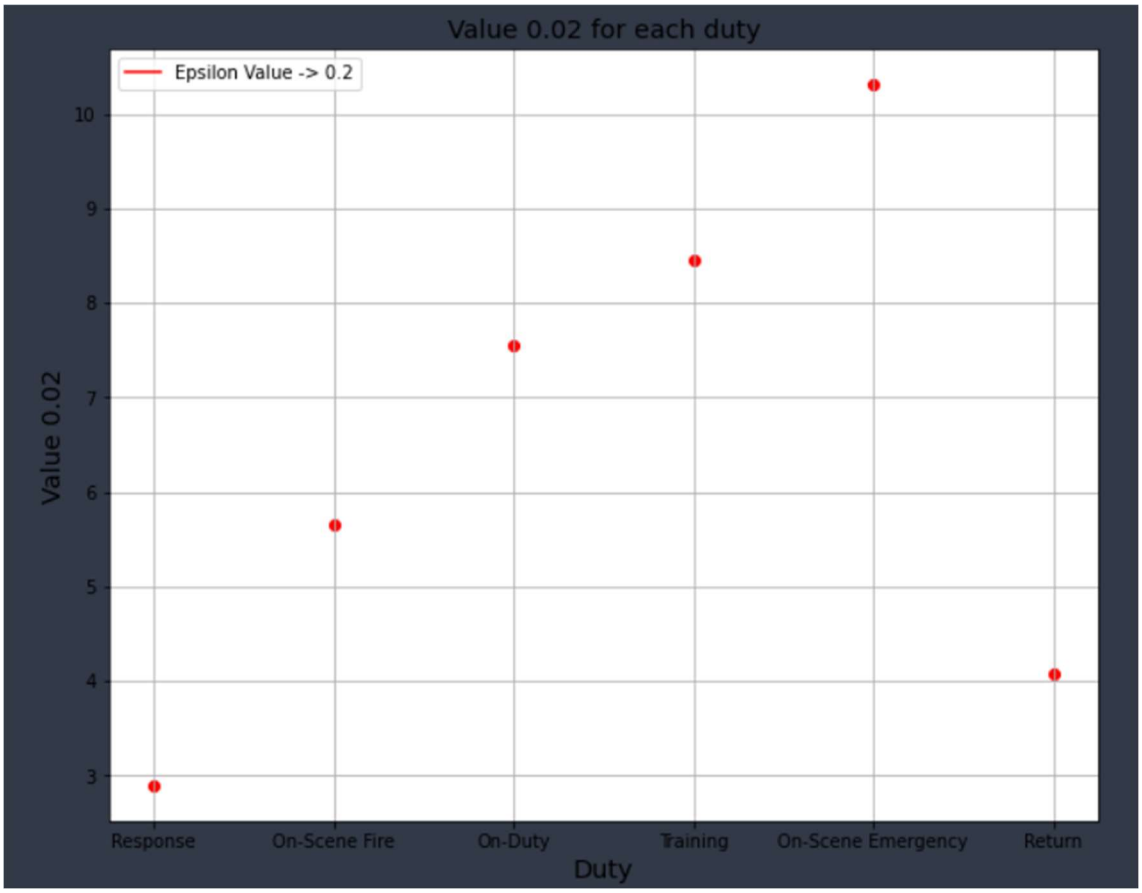

Figure 13 - Differential Privacy for each duty with epsilon value 0.2

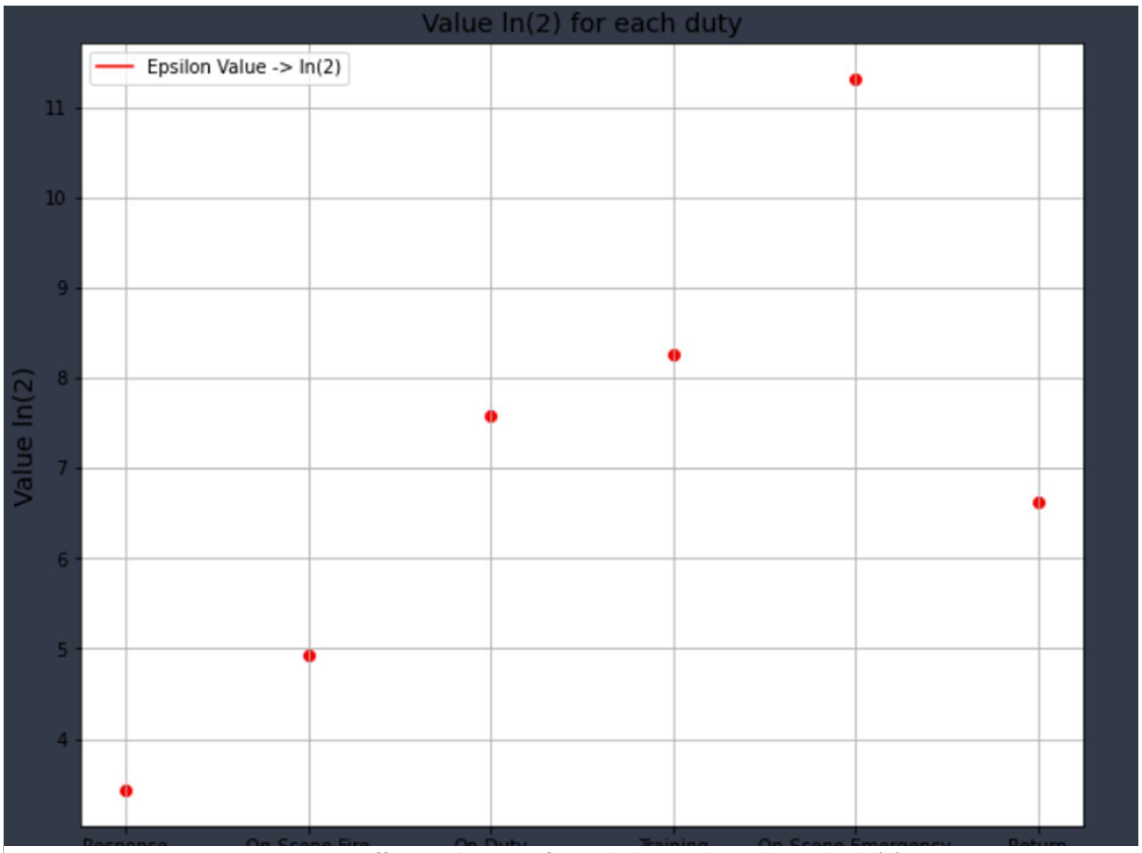

Figure 14 - Differential Privacy for each duty with epsilon value ln(2)

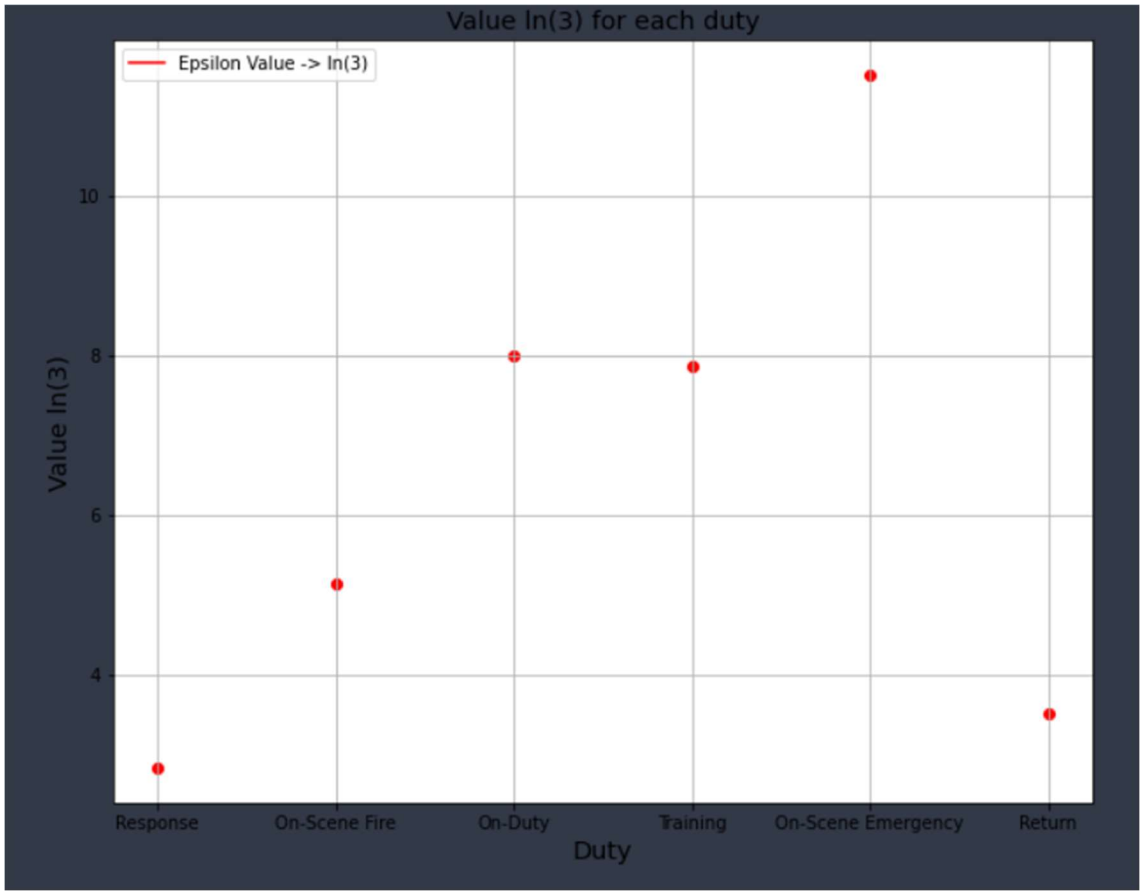

Figure 15 - Differential Privacy for each duty with epsilon value ln(3)

As we can see in the figure 9 the mean value of age and the results obtained for each epsilon value 0.01/0.2/ln(2)/ln(3) are low that's because the sensitivity, when we got more data to go through the lower is the sensitivity, which means the variance of the results are less noticeable, but in the figure 11 is the complete opposite of the figure 9.

In the figure 11 we analyze the mean age values for each duty and we can see that the epsilon values are very high, which means that data of that certain duty is low, and the variance applied to the mean age of that duty is really high.

We also can see the distributions of each value in figures that have graphs.

The conclusion that we retain is that bigger data analyses gets less sensitivity and smaller data gets more sensitivity. And all that affects the data. Comparing each one of the figures we can conclude that lower values of epsilon have good privacy values, but higher values of epsilon have not so good privacy.

Advantages and disadvantages

There are some advantages that we can conclude and these are:

- If the sensitivity is lower it doesn't change the data, and can't compromise the values.
- If we apply differential privacy, and high sensitivity, it's highly improbable that we can trace back to the original values.

And some of the disadvantages are:

- High cost, because it has to do several calculations, and if the dataset is big, more extensive will be the process.
- The second point of the advantages can also be a disadvantage because the data becomes useless.

References

Dataset Kaggle Firefighters: https://www.kaggle.com/fema/firefighter-fatalities

Software ARX Anonymization Tools