Athens University of Economics and Business

School of Business

Department of Management Science & Technology

Master of Science in Business Analytics

Program:	Full-time
Quarter:	2nd (Winter Quarter)
Course:	Business and Privacy Issues in Data Analysis
Assignment:	Data Anonymization Exercise
Students (Registration №):	Sakellaris Emmanouil (f2822215), Souflas Eleftherios-Efthymios (f2822217), Tsapatsaris Dimitrios (f2822218)

Table of Contents

Exercise A

1. Which attributes can act as quasi-identifiers and why?

Quasi-identifiers are pieces of information that are not themselves unique identifiers but are well correlated with an entity that when combined with its quasi-identifiers they can create a unique identifier. Quasi-identifiers can thus, when combined, become personally identifying information. This process is called re-identification (Wikimedia Foundation, 2022). As an example, Sweeney has shown that even though neither gender, birth dates nor postal codes uniquely identify an individual, the combination of all three is sufficient to identify 87% of individuals in the United States (Sweeney, 2000).

In the Public Use Microdata Sample (PUMS) dataset of Delaware State Census of 2010, there exist attributes like: Sex (gender), taking values Male or Female; Age, ranging values from 0 to 99 from which we can retrieve the birth year; Quarter of Birth, from which we can approximately get the birth month; Housing/Group Quarters (GQ) Unit Serial Number, from which we can get the housing quarter assigned within Delaware State that an individual lives in. Thus, apart from the full birth date (we only have the quarter of the exact year that people participating in the dataset were born), we almost have what Sweeney has proved to be sufficient to identify a large proportion of US citizens. In addition, in the dataset there also exist features of the race an individual belongs to, like Detailed Race Recode and Hispanic or Latino Origin; the sexual orientation by including indicators of marriage or unmarried partnership with Same-Sex Spouse; the Relationship to the Householder; the existence of Own or Related Child; and whether individual is living in a regular household, an institutional or a non-institutional Group Quarter. All forementioned attributes of a single record participating in the dataset can act as quasi-identifiers and when combined with other quasi-identifiers it is able to re-identify a unique person.

2. Which of the following properties holds for the data: they are anonymized, pseudonymized, or encrypted? Explain the key differences between the three approaches with respect to GDPR.

The data is pseudonymized. According to Article 4 of the General Data Protection Regulation, pseudonymization means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person (Official Journal of the European Union, 2016).

Encryption is a very old method. Plaintext is transformed to a ciphertext that only those that know the encryption secret can decrypt. Encryption is used mainly to ensure confidentiality. Encryption protocols can provide additional privacy enhancing mechanisms (Terrovitis, Privacy Protection in Information Systems, 2023).

Encryption and pseudonymisation belong to the category of personal data safeguards, according to Article 6 of GDPR. GDPR's principles of data protection are not applicable to anonymized information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable (Official Journal of the European Union, 2016).

3. Explain how a person can be identified.

Pseudonymized data is susceptible to link attacks with the use of publicly available datasets (e.g., voter lists, city directories) that can reveal the "hidden" identity (Terrovitis, Privacy Protection in Information Systems,

2023). Sweeney has linked medical data with voter lists by their common variables of ZIP Code, Birth Date and Sex to re-identify de-identified data as shown in Figure 1 (Sweeney, 2000).

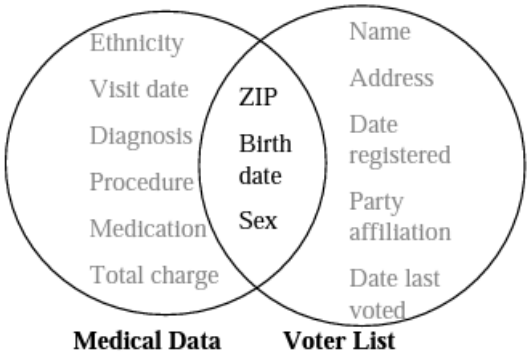

Figure 1 - Linking to re-identify data (Sweeney, 2000)

Also, if we know the exact or even approximate location of a person who we know that lived at 2010 in Delaware, we can find the Housing/Group Quarter Unit Serial Number that in the 2010 Census was allocated to Delaware Maps and then discover if he/she participates in the microdata sample dataset with the use of a subset of the forementioned quasi-identifiers. For example, if we knew a 70-year-old African American woman that in 2010 lived near St. Hedwig Roman Catholic Church in Wilmington, Delaware 19805 (US Census housing unit serial number for Delaware: 26, as found in the New Castle County PUMA Reference Map), we can deduce with a high probability that she participates in the 2010 Delaware Census Sample dataset for public use, especially because, in the 2010 Delaware Census, Persons of 65 years and over accounted for the 20.6% of the New Castle County (total) population, African American population accounted for 30.5% of the total population and female population accounted for 49.5% of the total population, as found in the New Castle County Census QuickFacts. Thus, we can reveal her hidden identity along with all other personal information that the specific row includes (79th row of the sample dataset– P0000026020100200020070030111010000000202002000).

4. Define differential privacy and explain the importance of the privacy parameter ε.

Differential privacy is a mathematical definition for the privacy loss associated with any data release drawn from a statistical database. The key idea is that the results of a query or an algorithm should be more or less the same whether a single individual's record participates in the input or not (Terrovitis, Differential Privacy, 2023). It is used to enable the collection, analysis, and sharing of a broad range of statistical estimates based on personal data, such as averages, contingency tables, and synthetic data, while protecting the privacy of the individuals in the data. Differential privacy is not a single tool, but rather a criterion, which many tools for analysing sensitive personal information have been devised to satisfy. It provides a mathematically provable guarantee of privacy protection against a wide range of privacy attacks, defined as attempts to learn private information specific to individuals from a data release. Privacy attacks include re-identification, record linkage, and differencing attacks, but may also include other attacks currently unknown or unforeseen (Wood, Altman, Bembenek, Bun, & Gaboardi, 2018).

As a mathematical definition, we say databases D_1 and D_2 differ in at most one element if one is a proper subset of the other and the larger database contains just one additional row. A randomized function K gives ε-differential privacy if for all datasets D_1 and D_2 differing on at most one element, and all S ⊆ Range(K),

$$Pr[K(D_1) \in \mathcal{S}] \leq e^{\varepsilon} \times Pr[K(D_2) \in \mathcal{S}]$$

The probability taken is over the coin tosses of K. Differential privacy is therefore an ad omnia guarantee. It is also a very strong guarantee since it is a statistical property about the behaviour of the mechanism and

therefore is independent of the computational power and auxiliary information available to the adversary/user. The forementioned Definition can be extended to group privacy as well (and to the case in which an individual contributes more than a single row to the database). A collection of c participants might be concerned that their collective data might leak information, even when a single participant's does not. Using this definition, we can bound the dilation of any probability by at most $e^{\varepsilon \times c}$, which may be tolerable for small c (Dwork, 2008).

A metric of privacy loss at a differential change in data (adding, removing 1 entry) or in other words what can be learned about an individual because of his/her private information being included in a differentially private analysis is limited and quantified by a privacy loss parameter, usually denoted as epsilon (ε). Privacy loss can grow as an individual's information is used in multiple analyses, but the increase is bounded as a known function of ε, and the number of analyses performed (Wood, Altman, Bembenek, Bun, & Gaboardi, 2018). We tend to think of ε as, say, 0.01, 0.1, or in some cases, $\ln 2$ or $\ln 3$. If the probability that some bad event will occur is very small, it might be tolerable to increase it by such factors as 2 or 3, while if the probability is already felt to be close to unacceptable, then an increase by a factor of $e^{0.01} \approx 1.01$ might be tolerable, while an increase of e, or even only $e^{0.1}$, would be intolerable (Dwork, 2008).

Exercise 2

1. Use the Amnesia anonymization tool to apply k-anonymity to the dataset. Comment on the resulting dataset.

Firstly, with the use of Python and Jupyter Notebook we separated the dataset into the underlying fields, created a table of the 89,924 rows for the person records only and extracted those records into a comma-separated-values (csv) file.

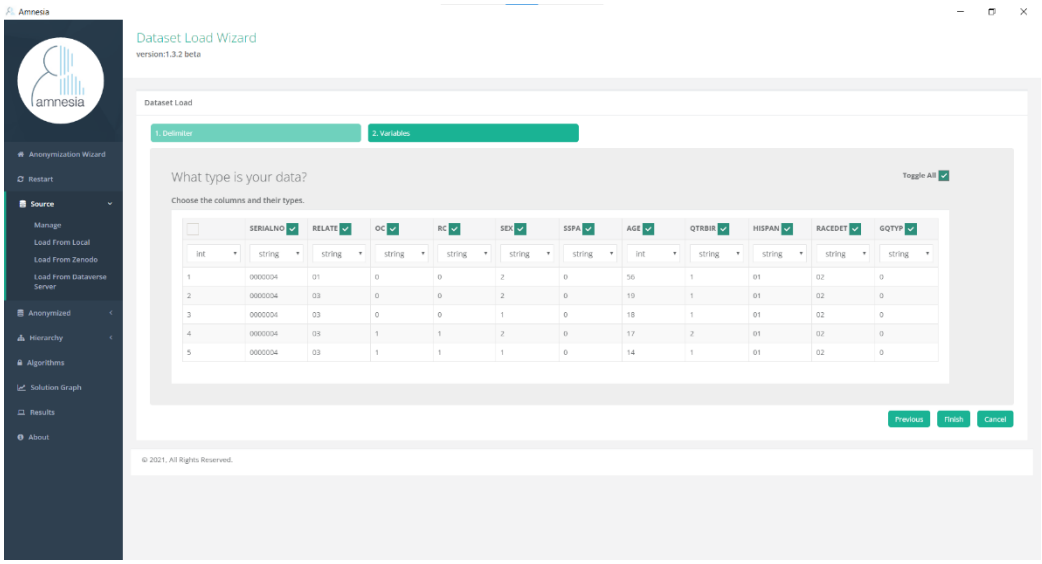

Figure 2 - Loading dataset to Amnesia

Then, we loaded the file to the Amnesia anonymization tool to apply k-anonymity to it, selecting only the fields that could potentially be used to identify a person i.e., the quasi-identifiers specified in the first exercise, as it is shown in Figure 2. In Figure 3, we can observe the percentage of k-anonymity the dataset included before the execution of the k-anonymity algorithm. After the creation of hierarchies for all quasi-identifiers, we executed a k=3 anonymity algorithm to the dataset. In Figure 4, we can observe the solutions graph provided by Amnesia. Blue nodes indicate safe solutions and red nodes unsafe.

We chose not to transform an unsafe solution (red node) to safe by applying suppression (eliminating rows) via the Amnesia tool, because Public Use Microdata Sample (PUMS) files contain records representing 10-percent samples of the occupied and vacant housing units in the United States and the people in the occupied units (group quarters people also are included), containing also individual weights for each person and housing unit, which when applied to the individual records, expand the sample to the relevant total (U.S. Census Bureau, 2014).

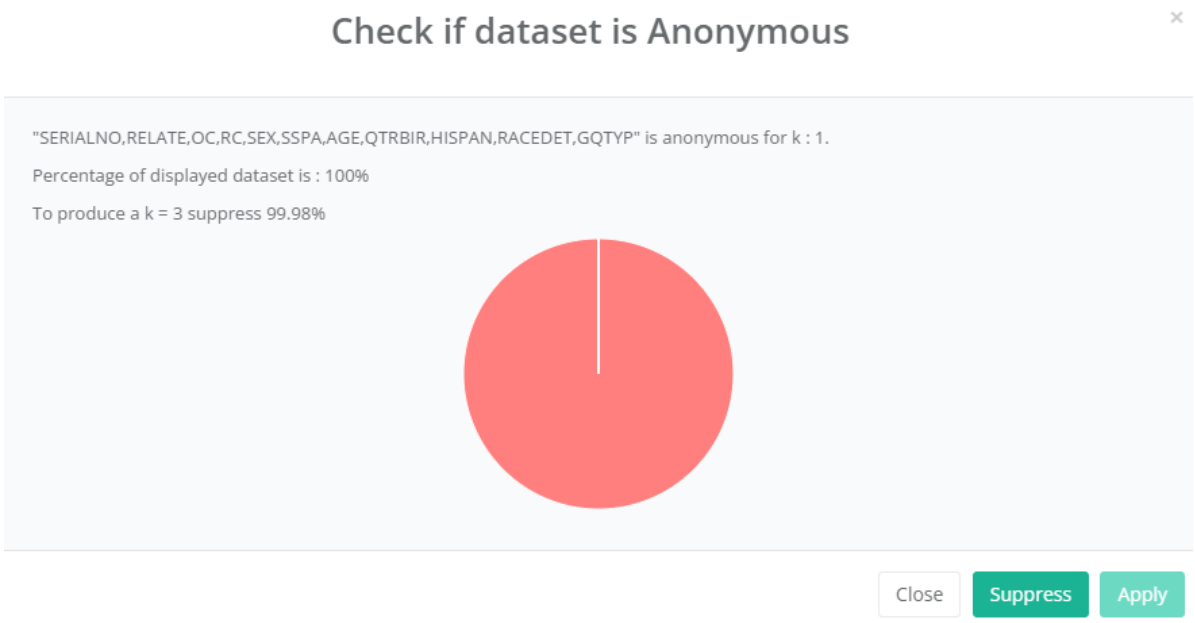

Figure 3 - Information on k=3 anonymity before applying algorithm

The solution provided by Amnesia included five quasi-identifiers which created a k=5 anonymity dataset, as shown in Figure 5. The resulting dataset includes the County that the house/group quarter each person lives belongs to, which was found from [US Census 2010 Delaware Maps], the Gender of the individuals (Males: 1, Females: 2), their Age in nine groups (0-6, 6-12, 12-18, 18-30, 30-45, 45-65, 65-75, 75-85, 85-99), and two indicators for having an own or related child respectively under 18 years old. A short overview of the resulting k-anonymized dataset can be observer in Figure 6.

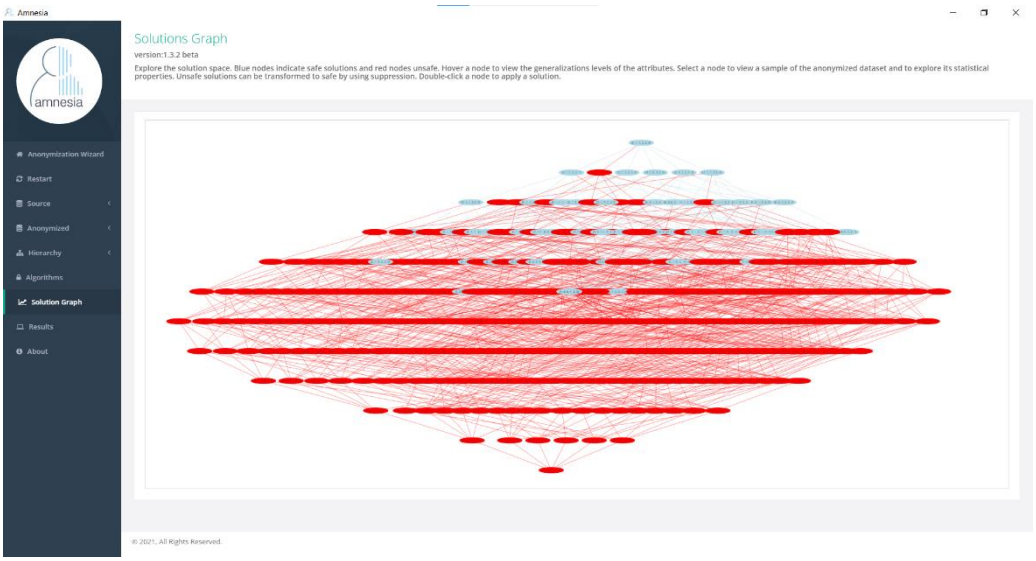

Figure 4 – Solutions Graph

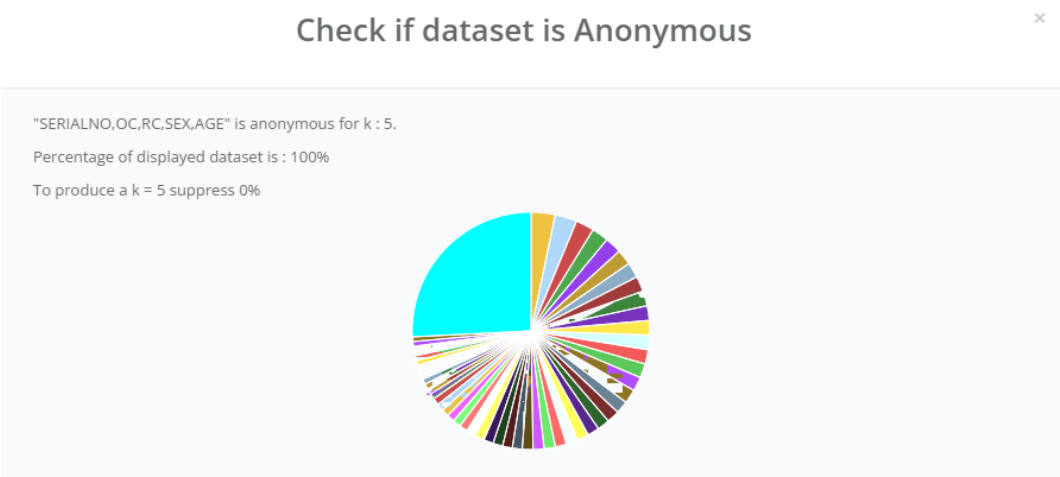

Figure 5 – Anonymous dataset with k=5

	SERIALNO	OC	RC	SEX	AGE
0	Sussex	0	0	1	75-85
1	West_Central_New_Castle	0	0	1	18-30
2	Kent	0	0	2	18-30
3	Sussex	0	0	1	75-85
4	Sussex	0	0	2	45-65
...	...	...	...	...	...
89919	Sussex	1	1	1	0-6
89920	Kent	0	0	2	18-30
89921	South_New_Castle	0	0	2	18-30
89922	Kent	0	0	2	45-65
89923	Kent	0	0	2	65-75

89924 rows × 5 columns

Figure 6 – Resulting Dataset

2. Plot the distribution of numeric features in the dataset using histograms.

The resulting dataset included only Age in groups as numeric feature, whose histogram can be observed below.

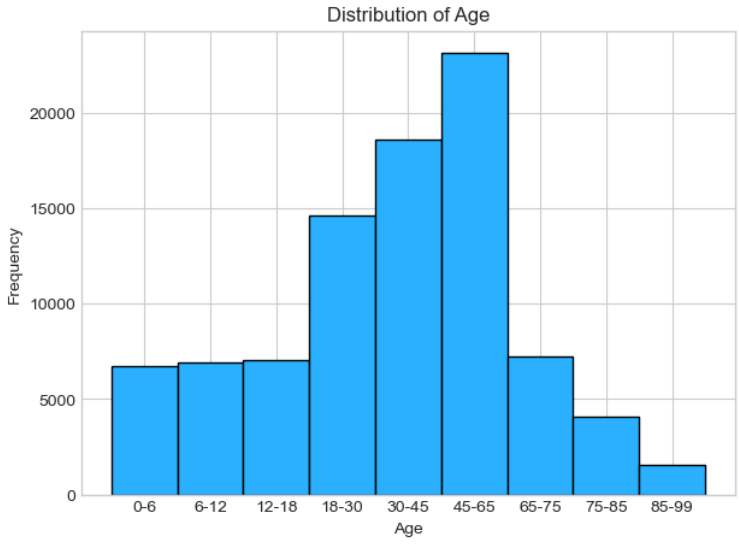

Figure 7 - Distribution of Age (histogram)

3. Apply a random noise mechanism to some of the numeric columns using the Gaussian mechanism. The noise should be added to the original values in a way that preserves differential privacy.

We selected to apply the Discrete Gaussian mechanism in differential privacy, as proposed by Canonne, Kamath and Steinke, re-purposed for approximate (ε, δ)-differential privacy (Canonne, Kamath, & Steinke, 2023). We let sensitivity parameter to be equal to 1 for parametrizing the amount of how much noise perturbation to be required in the mechanism. We set delta (δ) parameter close to 0, in order to apply an ε-differential privacy mechanism to the numeric features. As ε parameter we chose the value selected by the US Census Bureau for the 2020 US Census, which was the first Census to be applied differential privacy to it, for which an ε = 17.14 for statistics based on people (the "persons file") was chosen (Garfinkel, 2022). The numeric features we chose to apply the Discrete Gaussian mechanism are individuals' Age and Quarter of Birth.

4. Calculate the differentially private averages for the individuals using the noisy data.

The mean age of the ε-differential private feature for the individuals is 38.9, whereas the original was 38.3. The differentially private mean quarter of birth is 1.5, which is equal to the original mean quarter of birth. Thus, we can infer that for both features the collection, analysis, and sharing of statistical estimates based on personal data is enabled, while protecting the privacy of the individuals in the data.

5. Plot the distribution of numeric features after the noise addition. Try different values of the ε parameter. Comment on the effect of the differential privacy on the results.

We can observe in Figure 8 and Figure 9 that the distribution of Age is approximately the same although it is applied a differential privacy mechanism to it. However, this does not stand for the Quarter of Birth feature, as the first and fourth quarters have a huge decrease in their frequency and the second and third a huge increase.

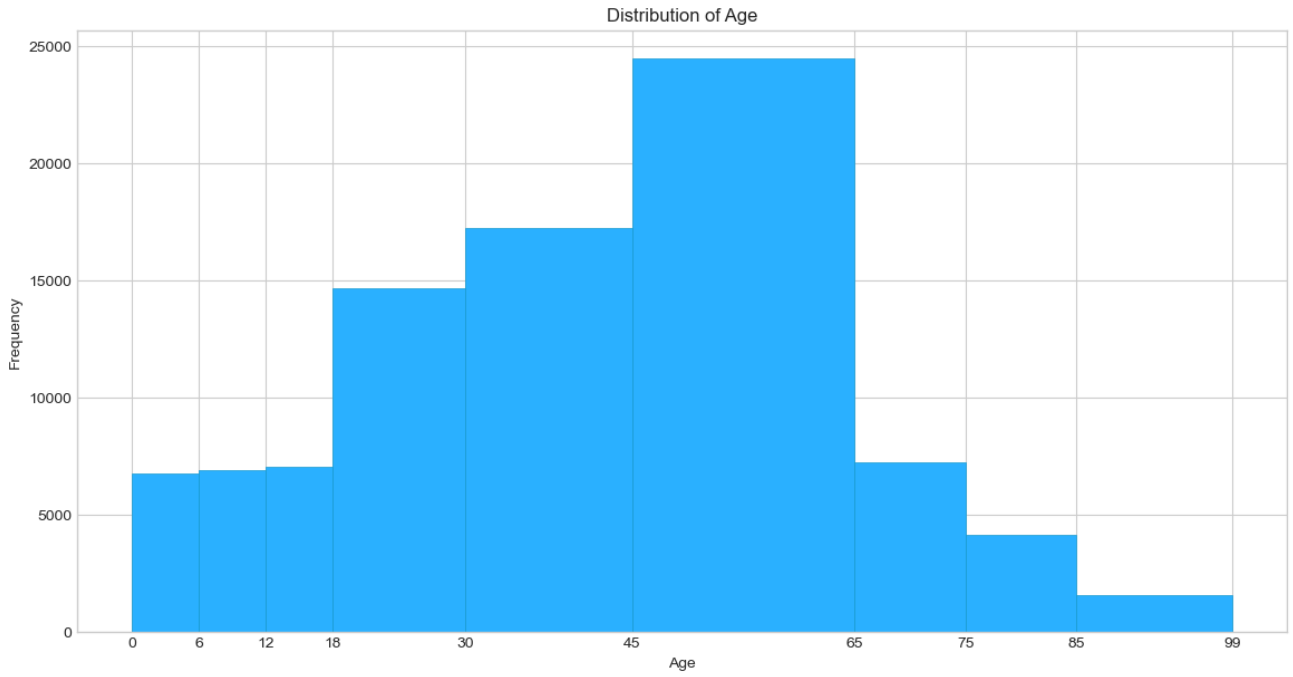

Figure 8 - Original Distribution of Age

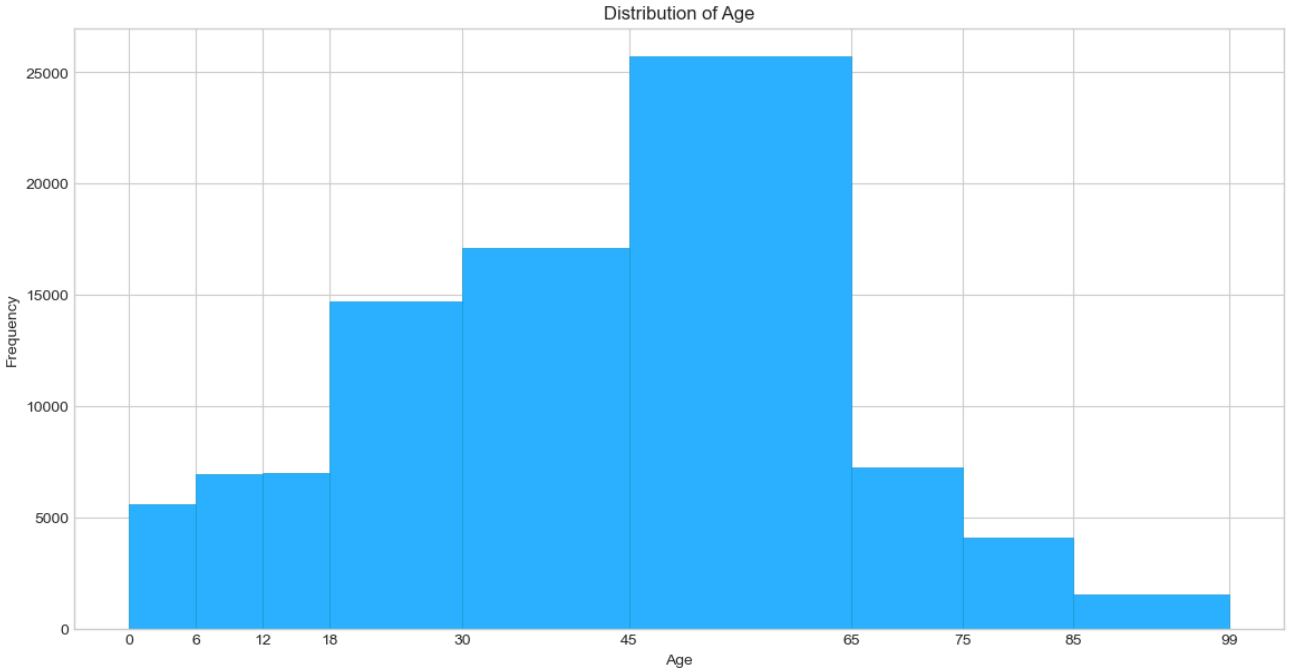

Figure 9 - Distribution of Age after noise addition

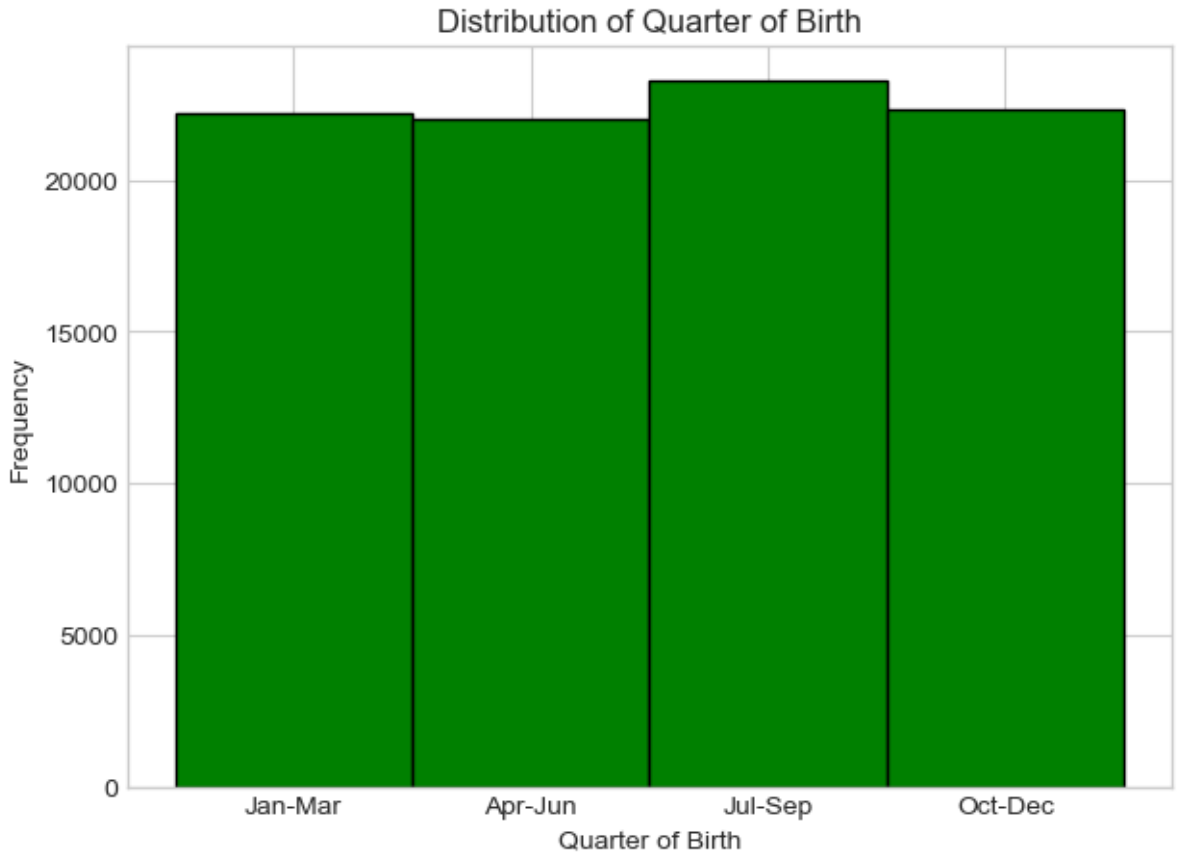

Figure 10 - Original Distribution of Quarter of Birth

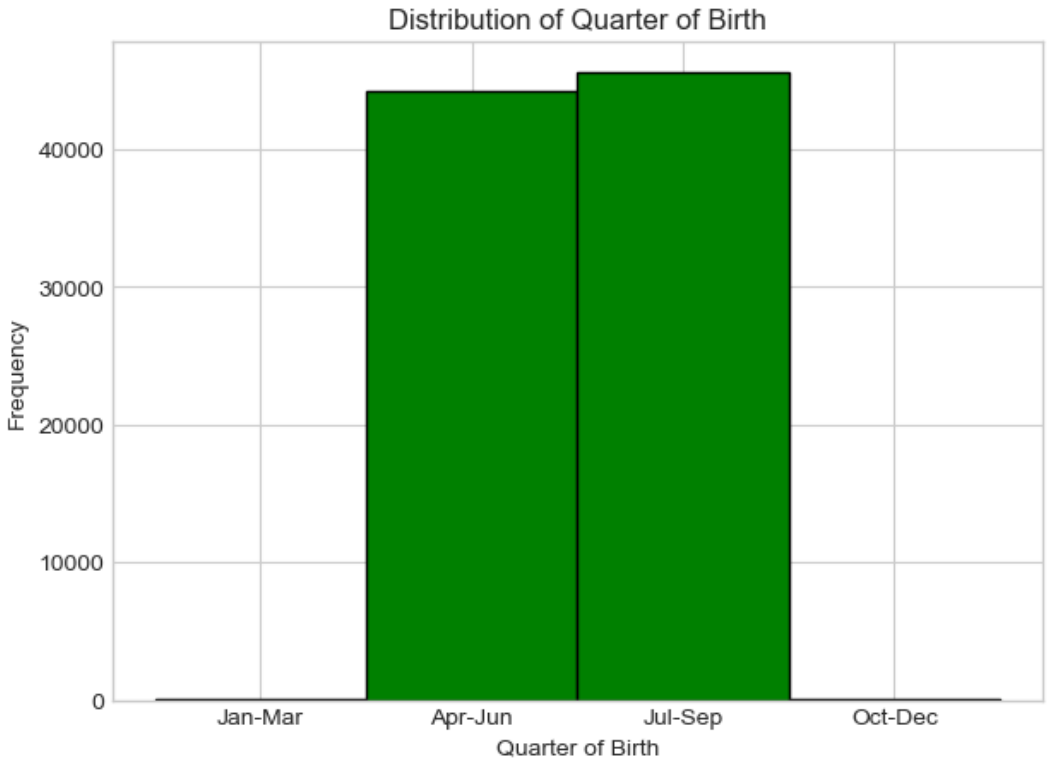

Figure 11 - Distribution of Quarter of Birth after addition of Discrete Gaussian noise

This is happening because we apply a Gaussian noise to a non-Gaussian discrete distribution. We then applied a Geometric noise, and it resulted in approximately the same distribution as its original, as shown in Figure 12.

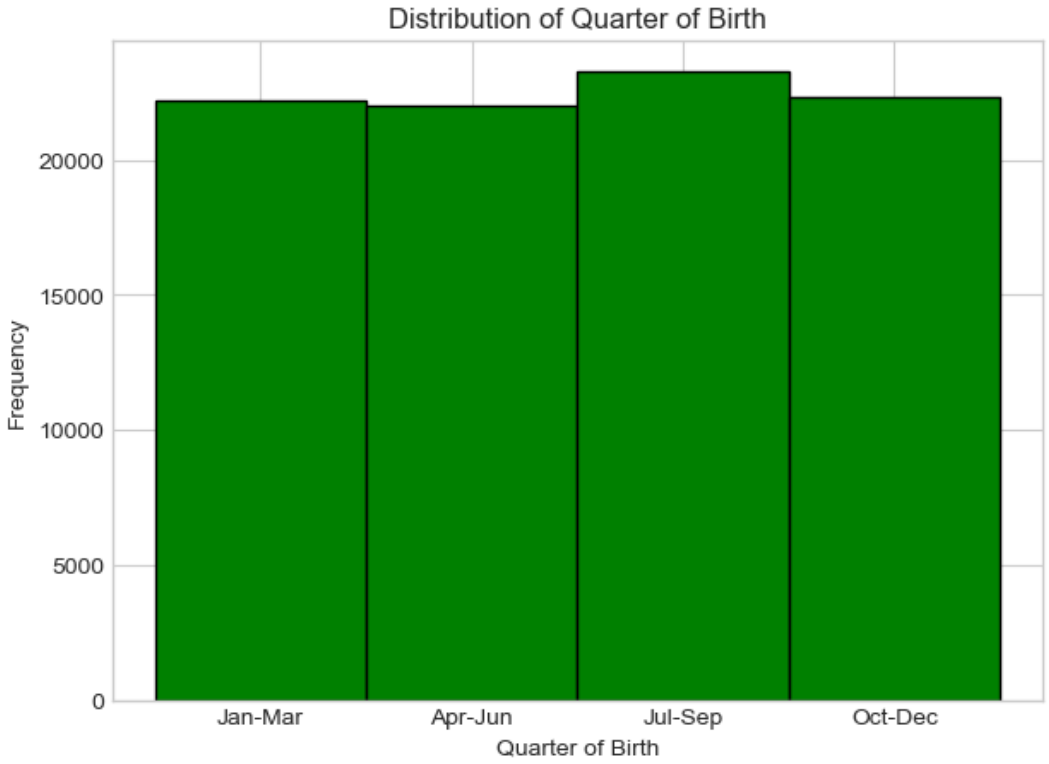

Figure 12 - Distribution of Quarter of Birth after addition of Geometric noise

Finally, we tried several different values for epsilon ranging from 0.01 to 100. We observed that as epsilon was increasing the distribution of the features were approximating the original distribution, thus also increasing the privacy loss.

References

Canonne, C., Kamath, G., & Steinke, T. (2023, May 8). The Discrete Gaussian for Differential Privacy. Retrieved from https://arxiv.org/pdf/2004.00010.pdf

Dwork, C. (2008). *Differential Privacy: A Survey of Results.* Microsoft.

Garfinkel, S. (2022). *Differential Privacy and the 2020 US Census.* MIT Case Studies in Social and Ethical Responsibilities of Computing. doi:https://doi.org/10.21428/2c646de5.7ec6ab93

Official Journal of the European Union. (2016, May 4). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC. Brussels: Official Journal of the European Union. Retrieved April 19, 2023, from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

Sweeney, L. (2000). Simple Demographics Often Identify People Uniquely. Pittsburgh, Pennsylvania, USA: Carnegie Mellon University. Retrieved May 2, 2023, from https://dataprivacylab.org/projects/identifiability/paper1.pdf

Terrovitis, M. (2023). Differential Privacy. *Slides from Lectures on Business and Privacy Issues in Data Analysis at AUEB MSc in Business Analytics*. Athens, Attica, Greece.

Terrovitis, M. (2023). Privacy Protection in Information Systems. *Slides from Lectures on Business and Privacy Issues in Data Analysis at AUEB MSc in Business Analytics*. Athens, Attica, Greece.

U.S. Census Bureau. (2014). *2010 Census of Population and Housing, United States Public Use Microdata Sample (PUMS): Technical Documentation.* Retrieved May 07, 2023, from https://www2.census.gov/programs-surveys/decennial/2010/technical-documentation/complete-tech-docs/us-pums/pumsus.pdf

Wikimedia Foundation. (2022, February 7). *Quasi-identifier*. Retrieved May 2, 2023, from Wikipedia: https://en.wikipedia.org/wiki/Quasi-identifier

Wood, A., Altman, M., Bembenek, A., Bun, M., & Gaboardi, M. (2018). Differential Privacy: A Primer for a Non-Technical Audience. *Vanderbilt Journal of Entertainment & Technology Law Vanderbilt Journal of Entertainment & Technology Law, 21*(1), 211-214.