BCSE318L – DATA PRIVACY – GEETHA MARY A

DIGITAL ASSIGNMENT 1
DEMONSTRATION OF ANONYMIZATION TOOLS

AATHIRA S

21BCE3680

Contents

I. Introduction

There is an increasingly high amount of data being collected today which presents both opportunities and challenges. While this large warehouse of data may be used to gain valuable insights through data analysis, the privacy and security of the individuals associated with these data records are paramount. Anonymization plays a crucial role in addressing and mitigating such concerns. [1]

Anonymization works by removing and modifying personally identifiable information (PII) from the datasets. This facilitates valuable research and collaboration, as well as the sharing of sensitive information, all while protecting the individuals' privacy. [1]

The following section II discusses anonymization and its subset k-anonymization in greater detail. Section III delineates various anonymization tools available today. The capabilities of ARX, a leading anonymization tool, is demonstrated in section IV. Finally we conclude our findings in section V.

II. Anonymization

Anonymization aims to limit the ability to link (or match) released data to other external information sources while still maintaining the useability and applicability of the released data. Apart from personally identifiable information, there exists other attributes which could be used along with external information for linkage attacks. Such attributes are known as quasi-identifiers. They include explicit identifiers such as name, phone number and address, as well as attributes whose combinations can be used to uniquely identify individuals such as ZIP code, birth date, and gender [2]. A goal of anonymity is to release person-specific data such that the ability to link to other information using quasi-identifier is limited.

K-anonymization

A common approach within anonymity is k-anonymity. By applying this technique, we can ensure that any combination of identifying attributes within a dataset appears for atleast k individuals, thus creating the effect of hiding the individual with a crowd and thereby making it difficult to link specific data points to one single individual. This 'crowd' is referred to as an *equivalence class*. [3]

K-anonymity achieves such protection primarily by using the following methods:

- Generalization, where specific values are replaced with broader ranges, or more generalized values. For example, the age 29 can be replaced with an age group range of [20,30).
- Suppression, where data points that could potentially lead to reidentification are removed. For example, if multiple people are born in the 20th century (1900s), but none in the same decade (example, 1990s), then their birth year data can be suppressed as 19**.

Table 1 illustrates k-anonymization of patients' medical data. Here the quasi-identifiers are {*Race, Birth, Gender, ZIP*} and k=2. Therefore, there exists 2 tuples for which the quasi-identifiers are the same.

Table 1: 2-anonymity in sample medical database [2]

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	M	02141	Short breath
t2	Black	1965	M	02141	Chest pain
t3	Black	1964	F	02138	Obesity
t4	Black	1964	F	02138	Chest pain
t5	White	1964	M	02138	Chest pain
t6	White	1964	M	02138	Chest pain
t7	White	1964	M	02138	Short breath

K-anonymization is not a perfect anonymization technique as it may be susceptible to attacks based on homogeneity within an equivalence class. Consider tuples *t5* and *t6* in *Table 1*. Both tuples belong to the same equivalence class and have the same medical problem. Hence, an attacker whose target individual is White, born in 1964, and lives in the ZIP area 02138 will have *chest pain*. [4]

III. Anonymization Tools

In order to handle and anonymize large datasets we require some mechanism for automation. For this purpose, we have a wide array of anonymization tools at our disposition. Some popular applications include ARX data anonymizer, Clover Dx's Data Anonymization Tool, Amnesia and µ-Argus.

ARX Data Anonymization Tool

The ARX Data Anonymization Tool is a comprehensive open source software for anonymizing structured individual-level health data. It has been continuously developed and extended with further functionalities over the past years and is freely available on the project website. Using anonymization or de-identification, data can be transformed in a semi-automatic manner to reduce risks to the privacy of individuals when data is processed for secondary purposes or disclosed to third parties [5].

The key features of ARX are as follows [6]:

- Transformation models: More than 10 different ways of transforming data, including various ways of generalization, top and bottom coding, suppression, sampling, and micro-aggression.
- Privacy models: ARX supports more than 10 different methods for quantifying and protecting privacy, including traditional syntactical models such as k-anonymity, statistical models for population uniqueness, and state-of-the-art semantic models including a game-theoretic approach and differential privacy.
- Utility models: The tool supports more than 10 different models for quantifying the utility of data during the anonymization process, including general-[purpose methods reflecting data and fidelity or changes to value distributions as well as application-specific models

In the following section we illustrate how to use ARX to perform k-anonymization.

IV. Demonstration

Data Used

The dataset train.csv [7] represents the passengers on the Titanic, and their details, including which of the passengers survived its sinking.

Table 2: Data dictionary for data set [7]

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
name	Passenger name	
sex	Sex	
age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin Number	
embarked	Port of Embarkation	C = Cherbough, Q= Queenstown, S = Southampton

Table 3: Classification of attributes in dataset [7]

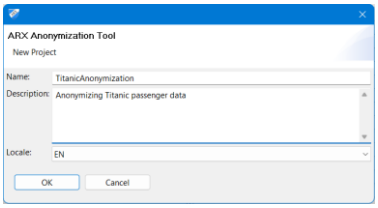
Variable	Classification
survival	Sensitive
pclass	Quasi-identifying
name	Identifying
sex	Quasi-identifying
age	Quasi-identifying
sibsp	Quasi-identifying
parch	Quasi-identifying
ticket	Identifying
fare	Quasi-identifying
cabin	Identifying
embarked	Quasi-identifying

Steps for anonymization

1. Create new project

On the taskbar ribbon at the top, go to File → New Project. Fill in the details and click *OK*.

Figure 1: Create new project on ARX



2. Import data

Again, go to the taskbar ribbon, File → Import Data. Select you file type. Here, we have used a CSV file. Click on *Browse* and select your file from the file explorer. Click on *Next*. Select *perform data cleansing* and the select *Next*. Confirm your data in the preview and *Finish*.

Figure 2: select input data's file type (left); Verify your data before completing import process (right).

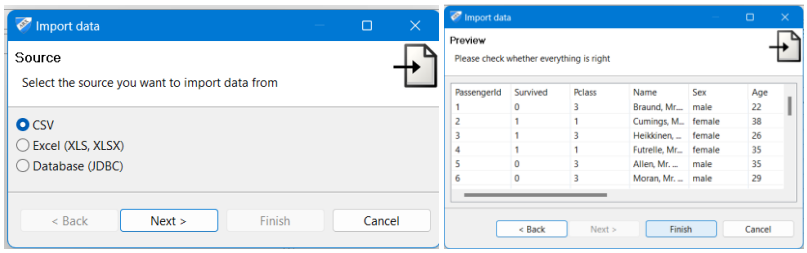
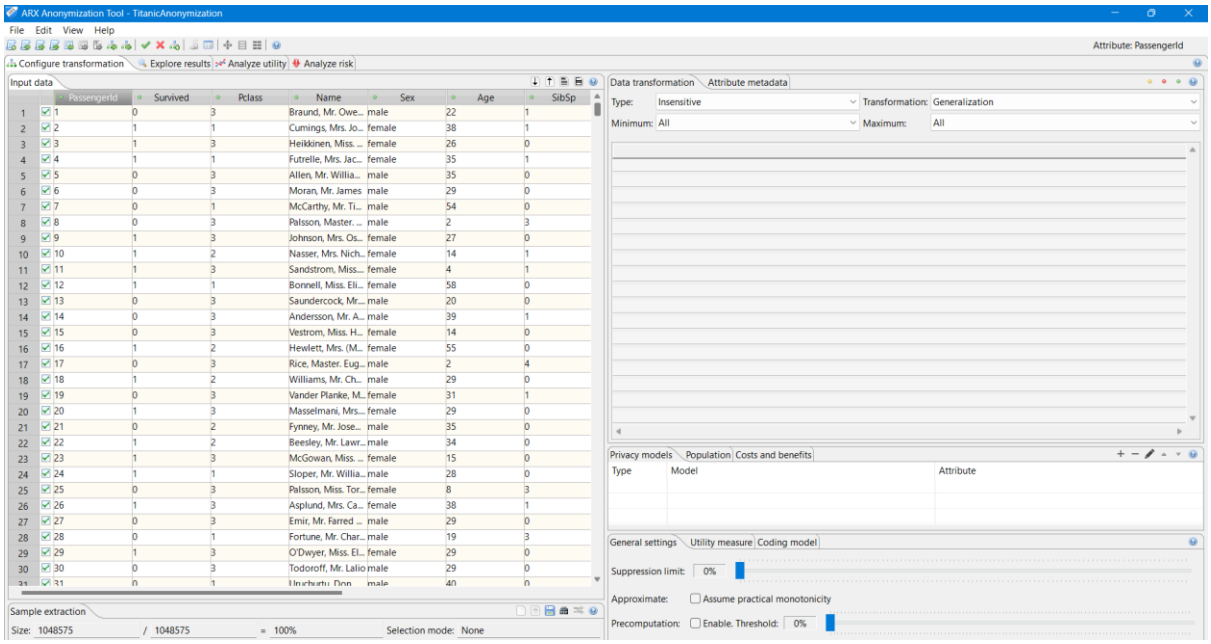


Figure 3: Data is imported into the project.

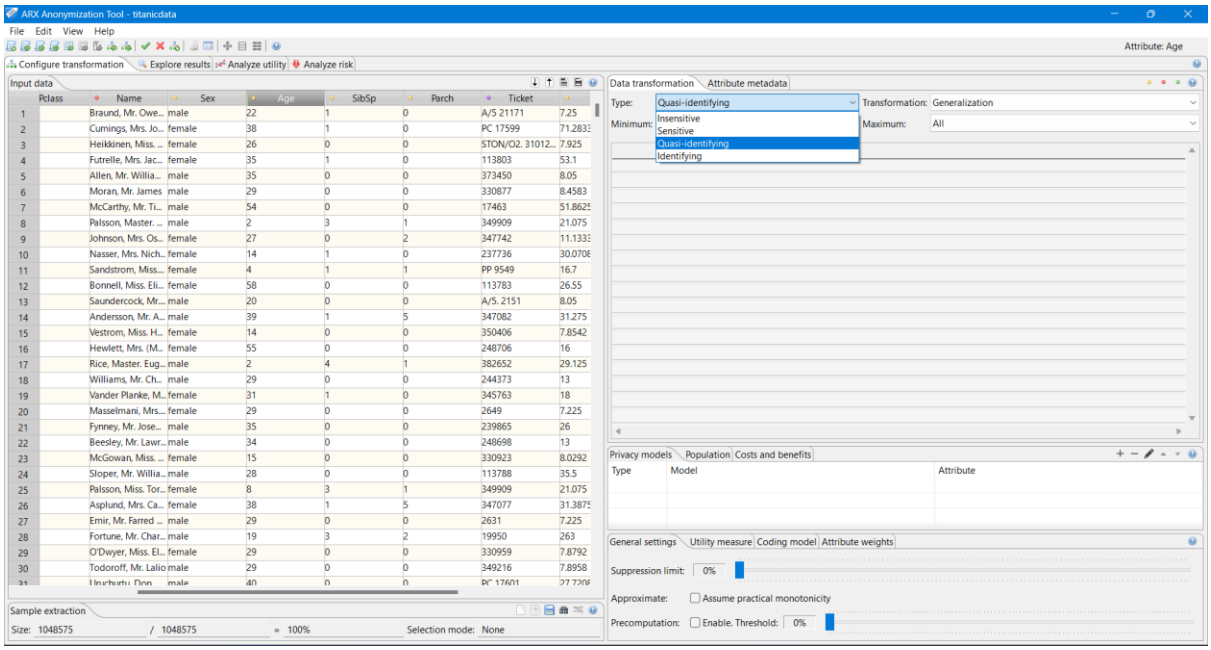


3. Set classification of attributes

For each attribute column in the dataset, set its type from one among the following:

- Insensitive
- Sensitive
- Quasi-identifying
- Identifying

Figure 4: Setting attribute type as 'Quasi-identifying' for attribute 'Age'



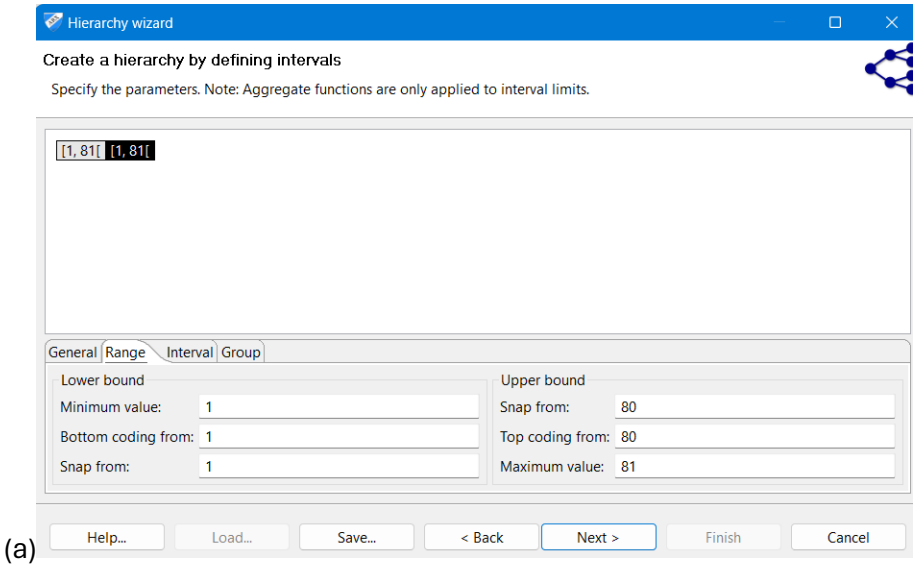
4. Set hierarchies for each attribute

Select the column for which you want to create a hierarchy. Now go to the taskbar ribbon select Edit→ Create Hierarchy. Now select the desired method to generate the hierarchy based on the column's data type.

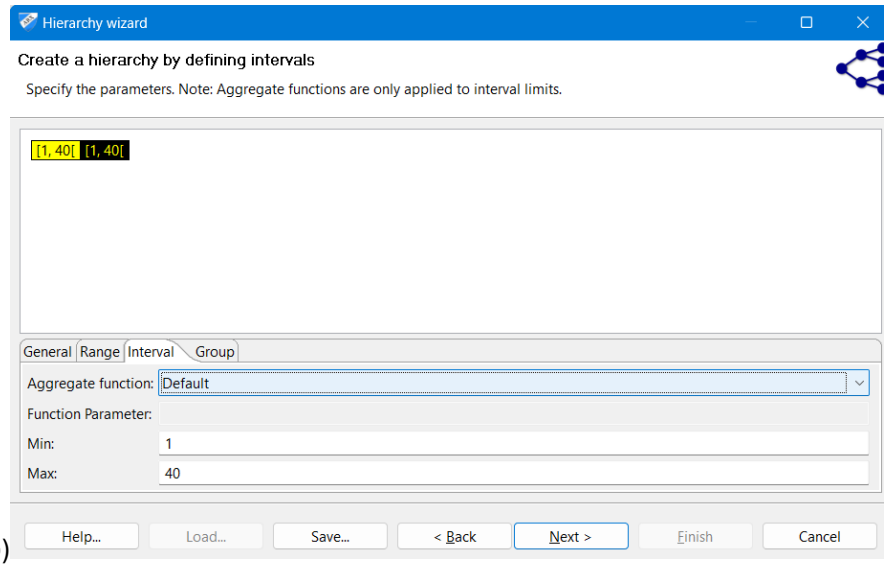
- Using dates, for date type data
- Using intervals, suitable for variables with a ratio scale.
- Using ordering, suitable for variables with ordinal scale.
- Using masking, suitable for alphanumeric strings
- Using priorities, suitable for a particular priority factor such as frequency.

Creating hierarchies using intervals, ordering and masking is illustrated in figures 5, 6 and 7 respectively.

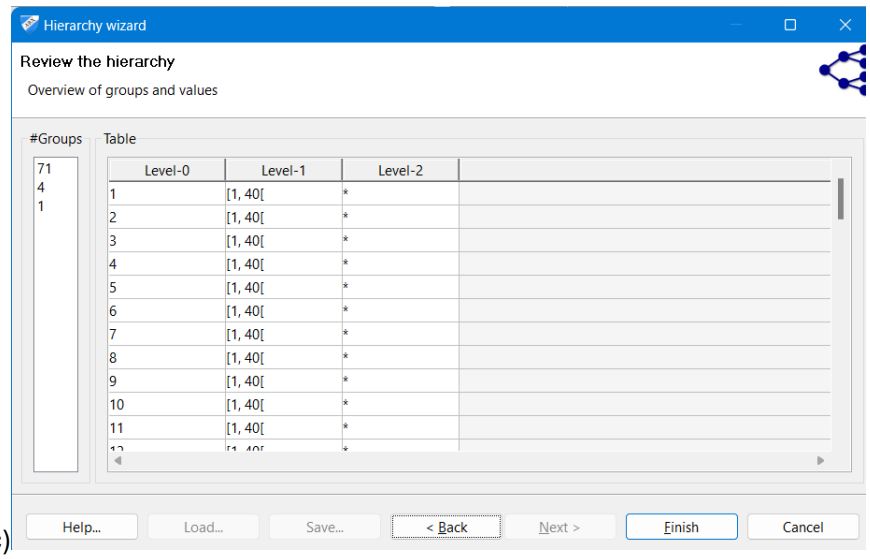
Figure 5: Creating hierarchy using intervals



(a)

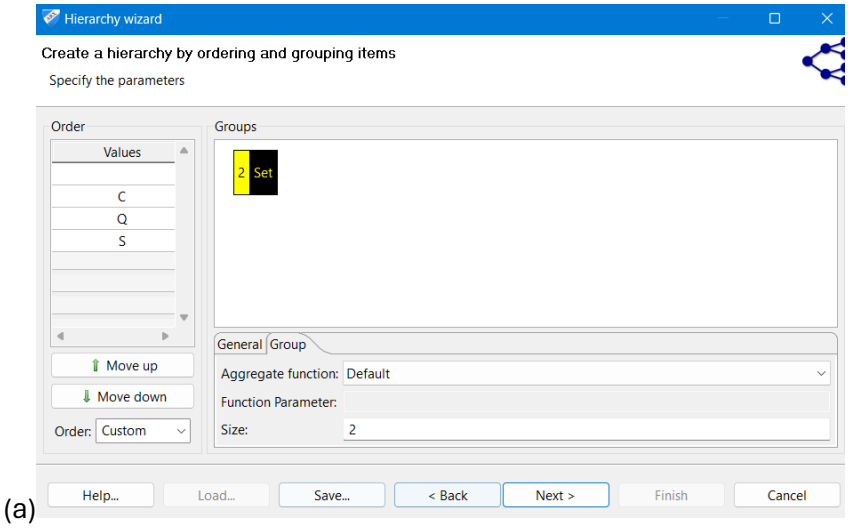


(b)

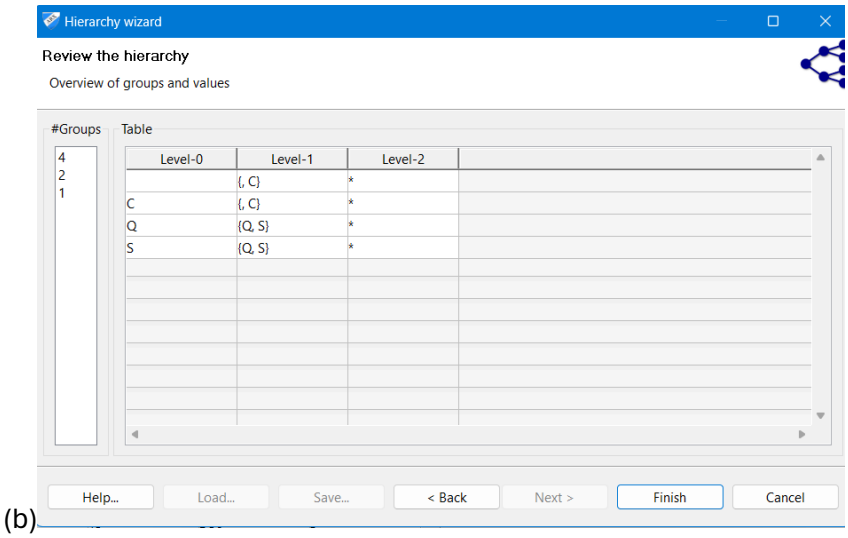


(c)

Figure 6: Create hierarchy using ordering

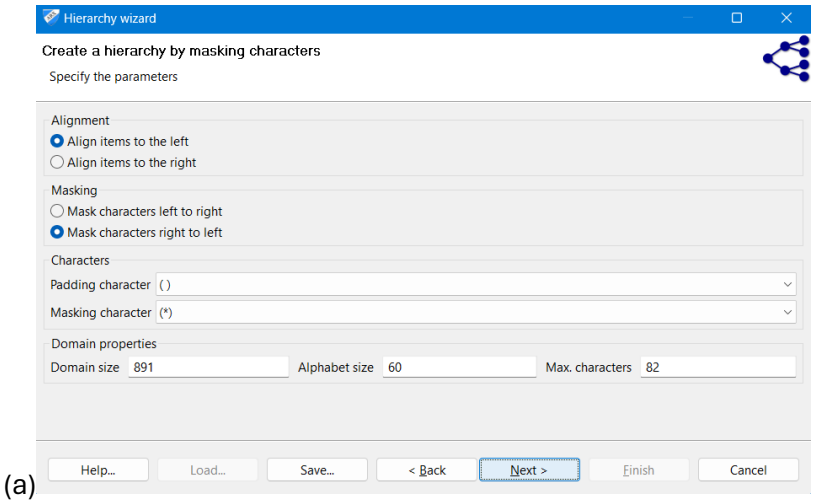


(a)



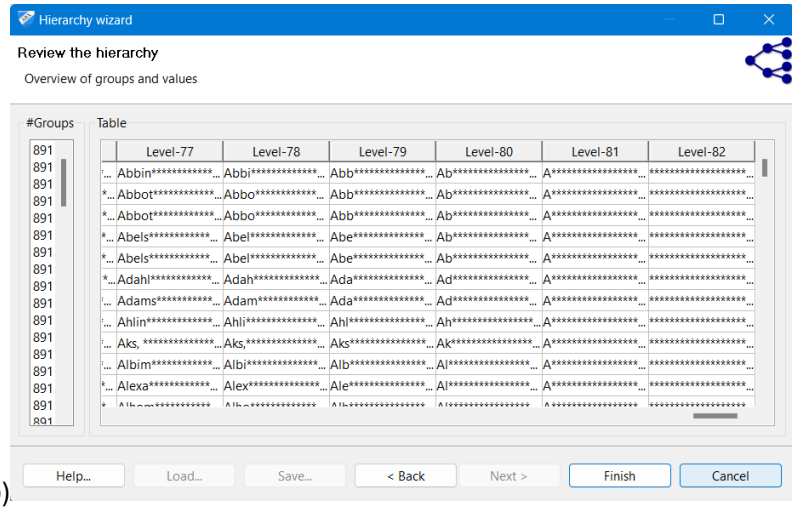
(b)

Figure 7: Creating hierarchy using masking



(a)

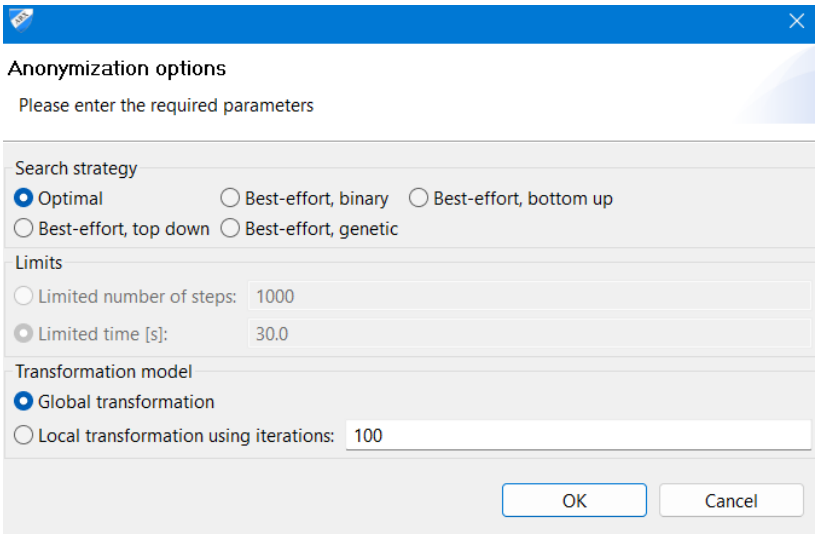
(b)



5. Anonymize the dataset

On the taskbar ribbon, select Edit→Anonymize. Choose the appropriate parameters and click *OK*.

Figure 8: Selecting parameters for anonymization



Results of anonymization

The results of anonymization and its effectives can be viewed and analysed by using the exploring the different tabs on the application interface. The results are illustrated in figures 9 to 13. We observe a decrease in overall risk of re-identification posed by the data. However there is also a drop in the utility values of the data. Moreover, the effectiveness of various attacker models against the dataset has been reduced significantly. Thus we can conclude that the results of anonymization helped protect the data from re-identification to a large extent.

Figure 9: Comparing the raw data before and after anonymization. Equivalence classes are highlighted.

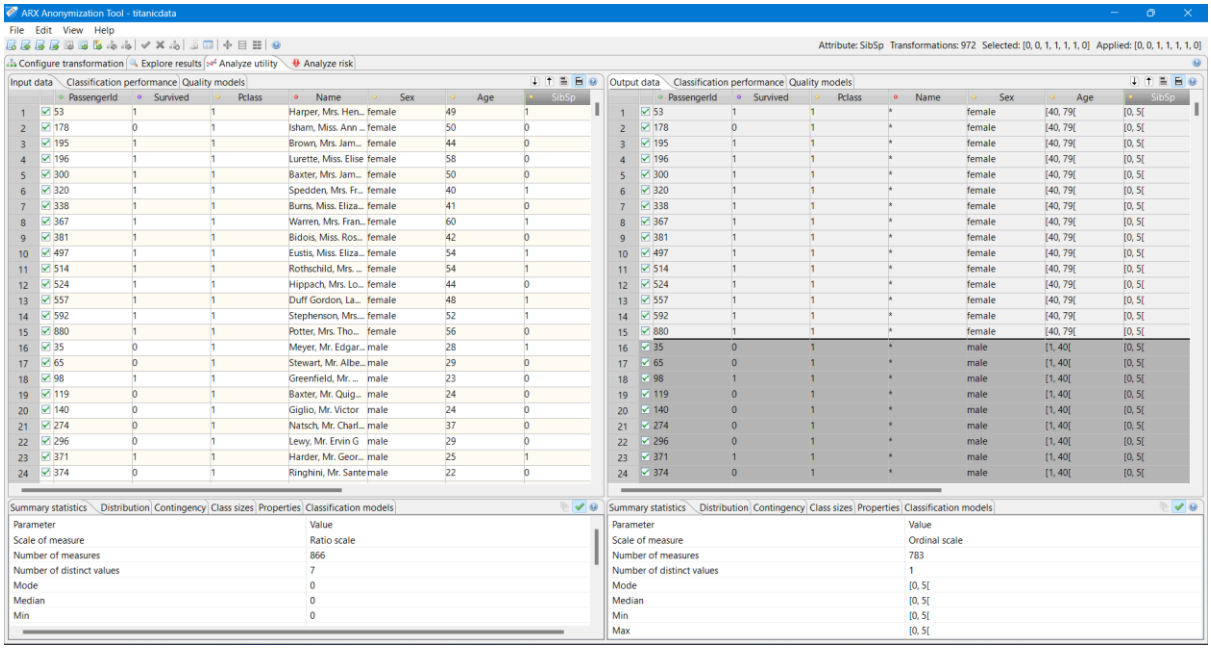


Figure 10: Comparing the quality of the data before and after anonymization, based on various attributes.

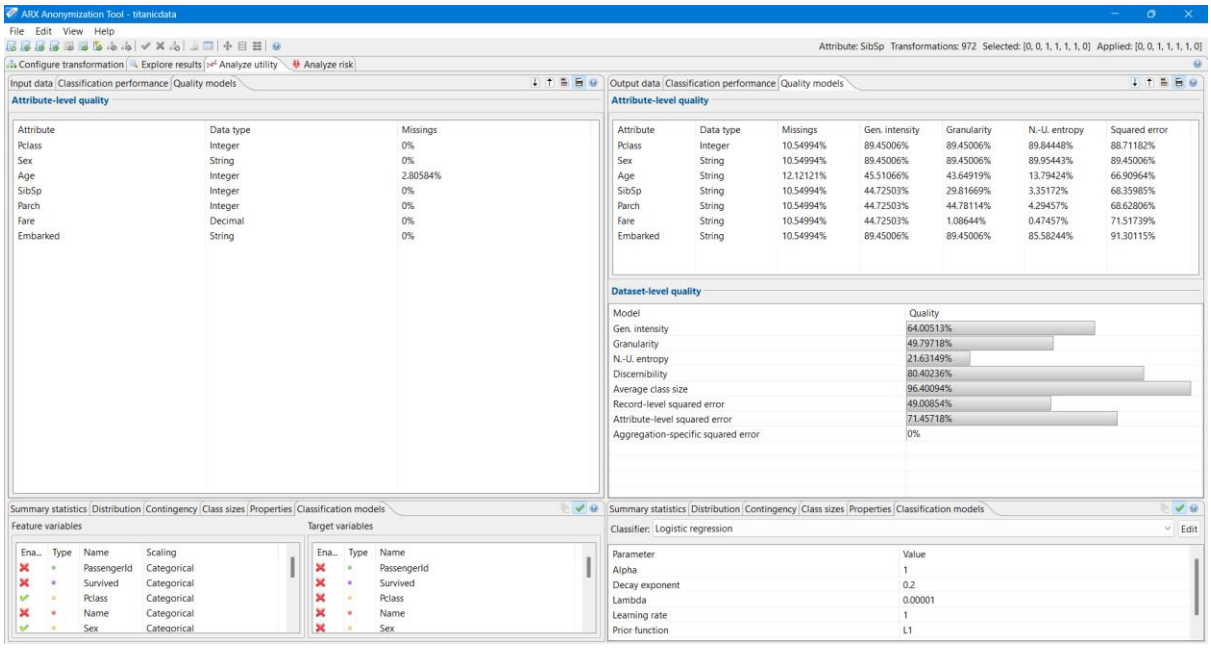


Figure 11: Comparing the distribution of risks across the data points before and after anonymization

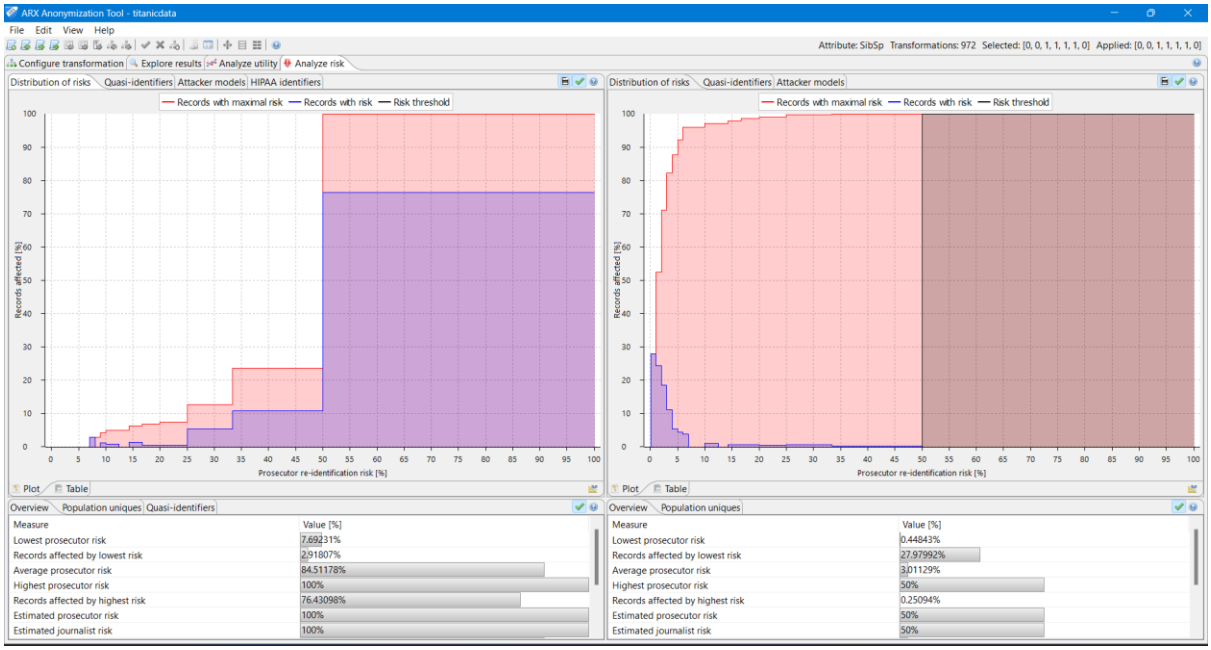


Figure 12: Comparing the effectiveness of attacker models before and after anonymization

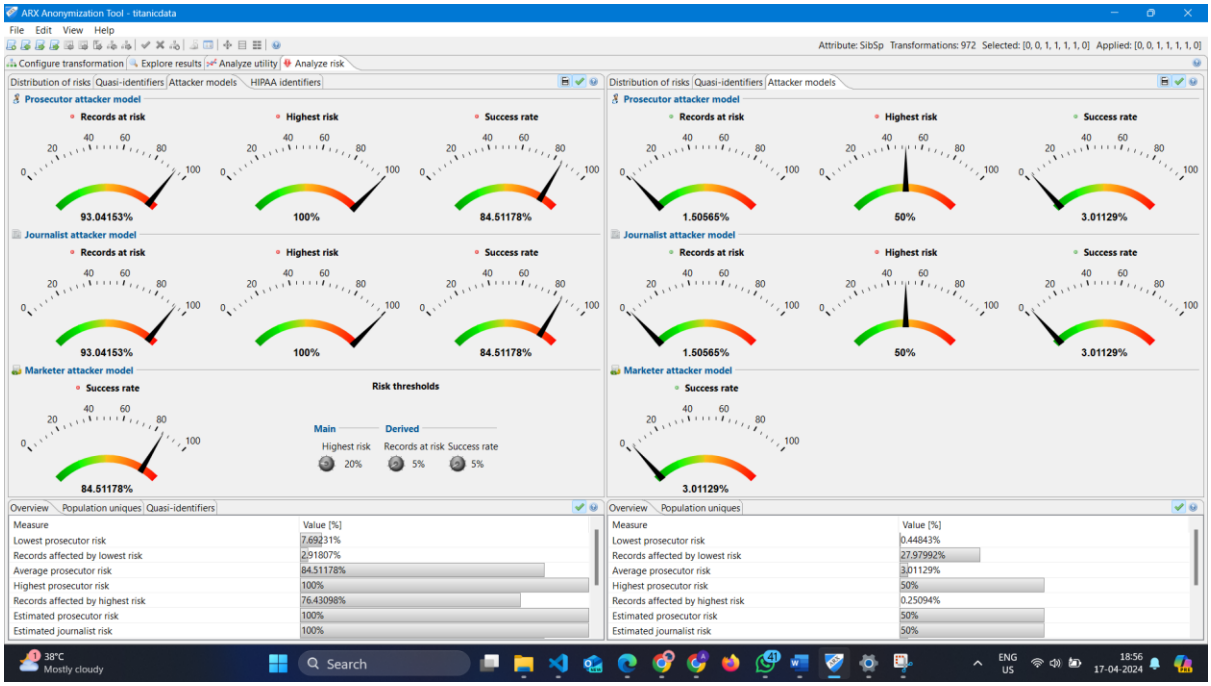
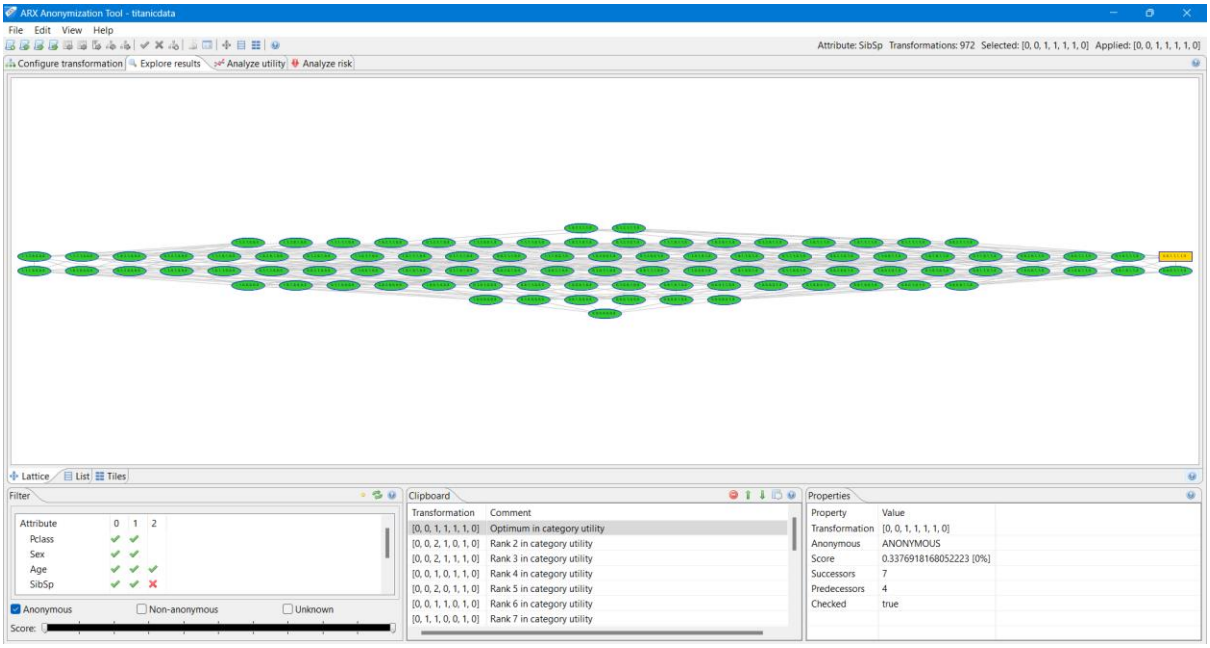


Figure 13: Exploring the results as nodes in a lattice of data in the anonymized dataset



V. Conclusion

This report has demonstrated the capabilities of ARX as a valuable tool for anonymizing data. Through its diverse anonymization techniques, ARX empowers users to achieve a strong balance between data privacy and analytical utility.

The demonstration showcased how ARX can be leveraged to achieve k-anonymity, a widely used technique for protecting individual privacy within datasets. By offering various methods like generalization and suppression, ARX allows users to tailor the anonymization process to their specific needs.

VI. References

[1] A. Chia, “Data Anonymization Explained: What You Need To Know,” 25 May 2023. [Online]. Available: https://www.splunk.com/en_us/blog/learn/data-anonymization.html.

[2] L. Sweeney, “ Achieving k-anonymity privacy protection using generalization and suppression,” *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,* pp. 571-588, 2002.

[3] L. Sweeney, “k-Anonymity: a model for protecting privacy,” *International Journal of Uncertainity, Fuzziness and Knowledge-Based Systems,* 2002.

[4] A. S. Nataraj Venkataramanan, Data Privacy: Principles and Practices, Taylor and Francis.

[5] ARX Deidentifier, “Overview of supported anonymization methods,” 2024. [Online].

Available: https://arx.deidentifier.org/overview/.

[6] ARX - Data Anonymixation Tool, “Overview of Features of the ARX Data Anonymization Tool,” 28 March 2021. [Online]. Available: https://www.youtube.com/watch?v=mkfJ-td-B94&ab_channel=ARX-DataAnonymizationTool.

[7] W. Cukierski, “Titanic - Machine Learning from Disaster,” Kaggle, 2012.