Privacy Preserving Data Anonymisation: an experimental examination of customer data for POPI compliance in South Africa

Nirvashnee Chetty (CHTNIR001)

Dissertation presented for the degree of

Master of Science

Department of Computer Science
University of Cape Town

Supervisor: Dr Andrew Hutchison

December 2019

PLAGIARISM DECLARATION

UNIVERSITY OF CAPE TOWN

DECLARATION

I, Nirvashnee Chetty, hereby declare that the work on which this dissertation/thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: Original Signed by Signed by candidate **Date:** 8 December 2019

ABSTRACT

Data has become an essential commodity in this day and age. Organisations want to share the massive amounts of data that they collect as a way to leverage and grow their businesses. On the other hand, the need to maintain privacy is critical in order to avoid the release of sensitive information. This has been shown to be a constant challenge, namely the trade-off between preserving privacy and data utility [1].

This study performs an evaluation of privacy models together with their relevant tools and techniques to ascertain whether data can be anonymised in such a way that it can be in compliance with the Protection of Personal Information (POPI) Act and preserve the privacy of individuals. The results of this research should provide a practical solution for organisations in South Africa to adequately anonymise customer data to ensure POPI Act compliance with the use of a software tool. An experimental environment was setup with the ARX de-identification tool as the tool of choice to implement the privacy models. Two privacy models, namely k-anonymity and l-diversity, were tested on a publicly available data set. Data quality models as well as privacy risk measures were implemented.

The results of the study showed that when taking both data utility and privacy risks into consideration, neither privacy model was the clear winner. The K-anonymity privacy model was a better choice for data utility, whereas the l-diversity privacy model was a better choice for privacy preservation by reducing re-identification risks. Therefore, in relation to the aim of the study which is to compare the results of data anonymisation to ensure that data privacy needs are met more than data utility, the result showed that the l-diversity privacy model was the preferred model.

Finally, considering that the POPI Act is still awaiting the final step to be promulgated, there is time to conduct further experiments in the various ways to practically implement and apply data anonymisation techniques in the day-to-day processing of data and information in South Africa.

ACKNOWLEDGEMENTS

First and foremost, I thank and praise God almighty for giving me the strength and the ability to undertake this research. Without His blessings on my journey, none of this would have been possible.

Throughout the writing of this dissertation I have received a great deal of support and assistance. I am grateful to my supervisor, Dr Andrew Hutchison, whose expertise was invaluable in the formulation of the research topic and keeping me on track to finish. You provided me with the advice that I required to choose the right direction and successfully complete my dissertation. Without your guidance and persistent help on this dissertation I would not have completed this paper. I thank you very much!

I would like to acknowledge my colleagues at the numerous companies I have worked for over the years, Justus Ortlepp, Gloria Silinda, and Lydia Molefe, just to name a few. You have been a great support and were always willing to help, affording me the leave to study when I needed it most.

In addition, I would like to thank my parents, especially my mother, for their wise counsel and sympathetic ear. You are always there for me when I need you and I appreciate your guidance in nurturing me to make my education my focus.

Last, but not least, to my wonderful husband, Sagren, who supported me from my first registration almost 17 years ago, to culminate here with the submission of my master's dissertation. You have always provided me with support, understanding, and the time necessary to complete my dissertation, especially since our daughter Elliana was born. She has been our biggest achievement to date, and I sincerely hope that my persistence and tenacity in reaching this milestone serves as an inspiration to her in that she can do anything she puts her mind to.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

Abbreviation	Term
AECS	Average Equivalence Class Size
API	Application Programming Interface
ASCII	American Standard Code for Information Interchange
CAT	Cornell Anonymisation Tool
CSV	Comma Separated Value
DM	Discernibility Measure
EU	European Union
FTC	Federal Trade Commission
GDPR	General Data Protection Regulation
GIC	Group Insurance Commission
GUI	Graphical User Interface
HIPAA	Health Insurance Portability and Accountability Act
I/O	Input Output
IL	Information Loss
OECD	Organisation for Economic Cooperation and Development
OLTP	Online Transaction Processing
PARAT	Privacy Analytics Risks Assessment Tool
PGP	Personal Genome Project
PHI	Protected Health Information
PII	Personally Identifiable Information
POPI	Protection of Personal Information
POPI	Protection of Personal Information Act
PPDM	Privacy Preserving Data Mining
PPDP	Privacy Preserving Data Publishing
SA	Sensitive Attribute
SSMS	SQL Server Management Studio
US	United States

1 INTRODUCTION

In this chapter, the introduction to the study is presented. The overview that provides the context to the study is presented first. Thereafter, the problem statement is identified. The research objectives in relation to the research aim are detailed resulting in the derived hypothesis and research questions that will be tested in this study.

1.1 Overview

Through the introduction of the Protection of Personal Information (POPI) Act of 2017 [2] (also known as POPIA), South Africa has taken its first steps towards enacting legislation that protects the privacy of individuals. With the enactment of the POPI Act, organisations and financial institutions in South Africa are required to protect customer information in the distribution and handling of data (both internally and externally) to ensure that they are compliant with the new regulations. In summary, the Act will impact all organisations and/or all parties within South Africa that are involved in the collection, processing, storage, and sharing of personal information. In the Act, 'personal information' also includes 'juristic persons' which are defined as legal entities whose data must be kept safe and the purposes of the use of this data be limited.

Businesses use information they store as an asset to give them a competitive advantage over their competitors by getting to know their customers better [3]. Advertisers also use it to find innovative ways to market their adverts to targeted consumers that are more likely to take up their products. However, the risk of releasing information, whether accidentally or on purpose, is always present. Therefore, the requirement for safeguarding personal information is of vital importance. As mentioned above, the sharing of personal information is a key concept that will be addressed by the POPI Act.

Data anonymisation has been found to be one of the most effective ways to protect the privacy of individuals [3]. Other data privacy legislation around the world has been in place for longer than the POPI Act; therefore, extensive work has already been done within the research community relating to medical data and the protection of health information [4]. The Health Insurance Portability and Accountability Act (HIPAA) of

1996 is one such legislation that specifically defines methods for de-identifying data through the HIPAA Privacy Rule [5] .

Protecting personal customer information is a relatively new concept in South Africa. There is no clear direction on what organisations must do to ensure that they are taking appropriate measures to protect their stored information. Therefore, in South Africa, further research needs to be undertaken in order to consider the protection of data that is stored and how this data is protected when being shared or disseminated.

1.2 Problem Statement

Whilst the usefulness of data that has been gathered by an organisation gives it a competitive advantage, it must be done in such a way that the privacy of an individual can still be protected. Many methods have been proposed to protect personal information over the years, for example cryptography, access control, and various other techniques [6]. However, these do not offer a guarantee of anonymity as the data is shared with recipients that are potentially unknown.

To address the need for anonymising data before distribution can occur, various techniques have been proposed [6]. As an overall key theme in how these techniques work, they aim to satisfy certain privacy objectives and ensure data is useful. Privacy models ensure that the objectives of protecting privacy are met and are enforced by privacy algorithms. These privacy algorithms transform data minimally but still ensure protection of the data.

Most of the use cases for the application of data anonymisation have been done within the area of healthy privacy, where data is transformed using generalisation and/or suppression. There is currently a lack of research related to the application of these data anonymisation techniques with regards to the POPI Act and is yet to be applied to this context.

However, in terms of the practical application of performing data anonymisation, software tools that implement well-known data anonymisation models are not readily

available. Software tools are indeed available, but these tools are not well documented or are complex to use. Furthermore, these tools have not been specifically applied to the context of data anonymisation for POPI Act compliance.

This study will evaluate software tools that have been created for data anonymisation for other data privacy regulations and apply these data anonymisation models within the software tool in order to ensure POPI Act compliance in the sharing of personal information.

1.3 Research Aim / Purpose

The main purpose of this research is to evaluate privacy models and data anonymisation tools and techniques for the practical application of anonymising customer data that is stored and used within the South African banking environment in such a way to ensure POPI Act compliance.

Two privacy models will be selected and applied to a customer data set. The outcome is intended to show the level of data utility achieved as well as measure re-identification risks in order to prevent personally identifiable information being leaked. The key consideration is whether further anonymisation being applied to the data set can result in lower re-identification risks. Furthermore, increasing the level of anonymisation could render the usability of the data set inadequate and, therefore, result in information loss and lower data quality. The trade-off between the two will be analysed in the results of this study, to show if there is a way to anonymise data appropriately to ensure POPI compliance whilst still retaining a sufficient level of utility.

1.4 Research Objective

The objective of this research is to:

1. Identify and evaluate existing data anonymisation techniques;
2. Identify a data anonymisation tool to practically perform the data anonymisation process;
3. Implement two privacy models within a data anonymisation tool;
4. Compare and contrast the results of the privacy models with one another with regards to the lowest level of privacy risk to ensure POPI Act adherence;
5. Compare and contrast the results of the privacy models with one another with regards to the highest level of utility retained so that the information is still useful; and
6. Establish the effects of privacy risks and data utility for static customer data in relating to POPI Act compliance.

The outcomes of this study will provide guidance to banking institutions on the way forward in the practical application of anonymisation tools to anonymise customer data when required to disclose personally identifiable information.

1.5 Hypothesis

In order to meet the objectives of the study, the following hypothesis has been formulated which will be tested during this study:

> Practically implementing the k-anonymity privacy model to anonymise customer data offers an effective solution to ensure POPI Act compliance by retaining the highest level of privacy without compromising the utility of the data.

1.6 Research Questions

In this study, the main research question specifically articulated in relation to the aim and purpose of the study that has been proposed follows:

1. Which privacy preserving anonymisation model is effective for anonymising static customer data for POPI Act compliance?

The following sub-research questions relate to the main research question:

1.1 Which privacy preserving data anonymisation technique is effective in practically anonymising static customer data for POPI Act compliance by ensuring the lowest level of privacy risk?

1.2 Can privacy-preserving data anonymisation techniques be practically applied to anonymise static customer data for POPI compliance in a way that preserves the greatest data utility?

1.3 What are the effects of the levels of privacy risk and data utility in ensuring POPI Act compliance of static customer data?

Table 1 below shows, in a summarised tabular view, how the research questions stated above map to the research objectives described in the previous section:

Table 1: Illustration of research objectives to research questions

Research Objective	Research Question
Identify and evaluate existing data anonymisation techniques	Which privacy preserving anonymisation model is effective for anonymising static customer data for POPI Act compliance?
Identify a data anonymisation tool to practically perform the data anonymisation process	Which privacy preserving anonymisation model is effective for anonymising static customer data for POPI Act compliance?

Implement two privacy models within a data anonymisation tool	Which privacy preserving anonymisation model is effective for anonymising static customer data for POPI Act compliance?
Compare and contrast the results of the privacy models with one another with regards to the lowest level of privacy risk to ensure POPI Act adherence;	Which privacy preserving data anonymisation technique is effective in practically anonymising static customer data for POPI Act compliance by ensuring the lowest level of privacy risk?
Compare and contrast the results of the privacy models with one another with regards to the highest level of utility retained so that the information is still useful	Can privacy-preserving data anonymisation techniques be practically applied to anonymise static customer data for POPI compliance in a way that preserves the greatest data utility?
Establish the effects of privacy risks and data utility for static customer data in relating to POPI Act compliance	What are the effects of the levels of privacy risk and data utility in ensuring POPI Act compliance of static customer data?

1.7 Research Methodology

To address the hypothesis and research questions in this study, an experiment was used. By performing an experiment, the results intend to show the cause-and-effect of using two privacy models for anonymisation when implementing these models on the same dataset. The results expect to show the researcher the data utility as well as data privacy outcomes. The data set obtained was from the Microsoft AdventureWorks 2016 sample data warehouse and this was used as it closely represented a realistic data set of static customer data that was publicly available. This dataset was imported into the chosen tool, ARX, where it was configured for anonymisation. Thereafter the privacy models were applied, and the results of each measured against utility and privacy.

1.8 Importance of the Research

The importance of this research is to provide a structured, clear way for organisations in South Africa to perform data anonymisation on customer data sets. This contribution is a tangible solution to protect personal information. The method in which the data anonymisation techniques and privacy models are implemented are practical in nature, thereby allowing the reader to make the clear link between theory surrounding data anonymisation and the practicality of using privacy preservation techniques. An important result of this research is that the techniques applied in this study can be used to anonymise customer data to ensure POPI compliance.

1.9 Limitations of the Study

This study did not take the following points into consideration and, therefore, are not in the scope of this study:

- The customer data set that is used in the study was specifically chosen as it contained a set of customer information that included data attributes that resembled real-world customer data. The results in this study relate to the customer data set selected;
- The customer data set used in this study represents a snapshot of data at a particular point in time which is referenced as static data and does not apply to streaming data that changes ad hoc; and
- Only two privacy models were selected to perform the data anonymisation techniques, but this was considered adequate for the study, within the defined scope.

1.10 Thesis Outline

In this study, the motivation and background information for data privacy is presented. This chapter sets the context and motivation for the study and the problem statement is discussed. A clear hypothesis with research questions is presented so that an overall aim can be achieved. In Chapter 2, the rationale and justification for data

privacy and data anonymisation is discussed. The relevant methods for data de-identification as well as legislation and laws surrounding data privacy is mentioned. In Chapter 3, a study of data anonymisation privacy model, privacy tools, and privacy techniques are performed. In Chapter 4, the approach and how the test environment was setup is outlined. In Chapter 5, the implementation of the selected tool is performed taking into considering the aspects of the design proposed. In Chapter 6, the results of the implementation of the test are documented. The evaluation of the results as presented in the previous chapter is done in Chapter 7. A summary of the results achieved, and findings is provided in Chapter 8, along with a conclusion relating to the initial aims and objectives of the study. Recommendations for future work is also presented in Chapter 8. Figure 1 below shows the graphical breakdown of the chapters in this study.

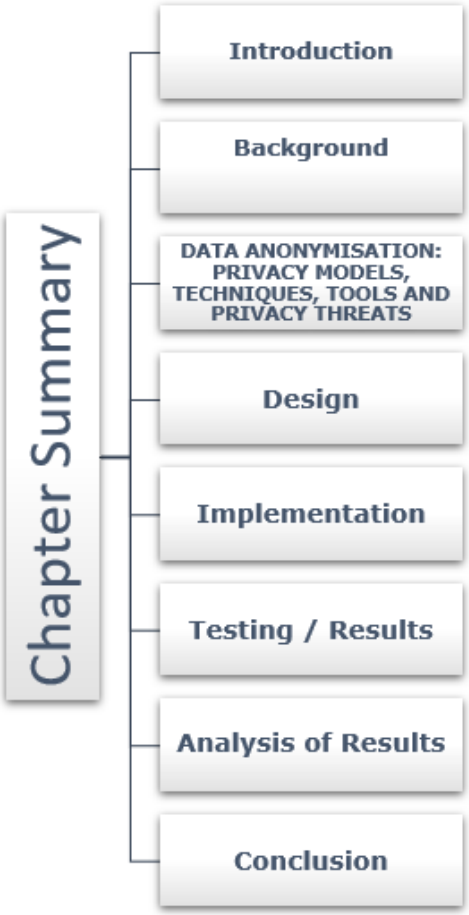


Figure 1: Thesis Outline

2 BACKGROUND

In this chapter, the background information related to data privacy is presented. The chapter starts with the initiatives that support the need for data privacy being detailed together with why data anonymisation was considered in this study as the way to protect customer information. Thereafter, terms and detailed definitions of data anonymisation and de-identification of data are provided. Lastly, the various legislations and laws surrounding data privacy are outlined with a description of data breaches that have occurred as a result.

2.1 Data Privacy Initiative

Recent developments with regards to data privacy in South Africa, notably the trends towards protecting the personal identifying information of customers, are accompanied by significant needs of privacy protection. In 2016, the United Nations Conference on Trade and Development proposed that an estimated 108 countries globally had enacted some sort of data privacy legislation [7].

As the United Nations have pointed out in this article [7], the challenges that have been faced by each of these countries have a common thread [7]:

(1) Long timeframes to pass legislation;
(2) Financial expenses to enforce the implementation of new data privacy legislation;
(3) A general lack of knowledge in the regulation of the new data privacy legislation as well as the limited cooperation between the public and private institutions.

A key point listed in the article above as a challenge is the lack of knowledge surrounding the implementation of the POPI Act in South Africa. As detailed in the study done more recently in 2017, referred to in article [8], in South Africa, a survey was performed to check the level of readiness of organisations to implement the POPI Act. An important result that was presented in the findings was that organisations that

took part in the research were not ready (at that stage) to be able to reliably ensure compliance of the POPI Act data privacy legislation.

If as a result of the POPI Act there is non-compliance by organisations in South Africa, there may be far-reaching consequences to them as a result. Penalties enforced for non-compliance can be quite costly, not to mention the reputational damage that can occur [9]. The Information Regulator has many options as a course of action for non-compliance: it may choose to prosecute the organisation by laying criminal charges against them together with the organisation also possibly being fined up R10 million. This could, however, be extended to include twelve months of prison time to be served as well. Prison terms can be extended up to ten years if a transgressor knowingly withholds information or obstructs the investigations into non-compliance. The Information Regulator may also issue an "enforcement notice" to immediately stop the storage and processing of personal information.

Consequently, in an effort to avoid the penalties that could potentially be faced for non-compliance, it is noted that there is clearly a need for data anonymisation approaches to be considered for implementation in preparation for the enactment of the POPI Act.

2.2 Why Perform Data Anonymisation?

Personally identifiable information (PII) is defined as information collected that in some way can identify the real identity of an individual or subject [10]. Additionally, the POPI Act refers to the term PII as well.

A few direct examples of PII taken from the study in [10] are:

- "Name (full name, maiden name, mother 's maiden name, or alias);
- A personal identification number (identity number, passport number, driver's license number, taxpayer identification number, bank account number, credit card number);
- Address information (street address, email address);
- Personal characteristics information (photographic images, fingerprints, handwriting, or other biometric data [retina scan, voice signature])."

To ensure the privacy of PII, safeguards need to be put into place to protect sensitive data. The POPI act specifically contains principles that govern how data should be stored, used and processed, and more specifically relates to the security of the sensitive data [2] . The principles are:

1. Accountability
2. Processing Limitation
3. Purpose Specification
4. Further Processing Limitation
5. Information Quality
6. Openness
7. Security Safeguards
8. Data Subject Participation

The principle, Security Safeguards and Controls, states that when personal information is collected, it must be sufficiently safeguarded from loss and being accessed unlawfully. The expectation is that sufficient measures (physically and electronically) must be taken by the organisation to ensure that the security and safeguard of the PII is in place [10].

A way of protecting PII is through data anonymisation. Data anonymisation involves removing the link between an individual and information that which identifies any individual's identity. An important consideration of doing this is to sanitise the information to a level that privacy and data utility is still maintained. By sanitising the data set, thereby removing personally identifiable information (PII), a comfortable level of anonymisation can be achieved [11].

A use case for data anonymisation within the banking industry and in context of this research is the anonymisation of customer data. Customer data is collected daily by banks and financial institutions. This stored data is considered 'personal information' as this data usually contains common personal information like names, date or birth and identity numbers; therefore, this information must be protected. In order to ensure the usefulness of data post anonymisation, the data must also be sufficiently anonymised, simultaneously ensuring minimal information loss [12].

Within the context of this study, the researcher aims to provide a practical implementation of anonymisation methods to sufficiently anonymise data and ensure minimal information loss.

2.3 Privacy Preserving Concepts

In modern day society, personal data or information is a commodity. There is constantly a need to release data external to an organisation, be it publicly or internally, so that it can be available to third parties for analysis without the disclosure of personal information that it contains [13]. There are many techniques and methods that try to address this problem. A few concepts are described below.

2.3.1 Privacy Preserving Data Mining (PPDM)

Firstly, data mining entails extracting patterns (knowledge) from data sets which can then be interpreted by an external user and represented in meaningful ways. Data transformation methods that cater for the extraction of knowledge from data whilst at the same time aiming to enforce rules to maintain privacy are known as privacy preserving data mining (PPDM) techniques[14] .

There are many approaches which have been adopted for privacy preserving data mining. These can be classified based on the following five dimensions as outlined in [15]:

1. **Data Distribution**

   This approach refers to the distribution of data. There are two types, horizontal distribution and vertical distribution. With vertical distribution of data, the values for different attributes reside different places whilst with the horizontal distribution of data, the different database records themselves reside in various places.

2. **Data Modification**

   Data modification is a way of changing the original values in the data set prior to release to a third party. Modification methods include blocking, perturbation, aggregation, sampling, and swapping.

3. **Data Mining Algorithm**

   Different data mining algorithms are available. These include classification data mining algorithms, like association rule mining algorithms, decision tree inducers, rough sets, Bayesian networks, and clustering algorithms.

4. **Data or Rule Hiding**

   This dimension refers to hiding data where it could be in the form of raw data or aggregated data.

5. **Privacy Preservation**

   Finally, the selective modification of data is the last dimension to be discussed, which is the most important. This technique ensures privacy preservation. Data is modified selectively so that a greater level of utility is obtained and done in a way that privacy can be preserved.

2.3.2 Privacy Preserving Data Publishing (PPDP)

One solution to realise privacy is to anonymise records in a data set before it can be published. Privacy Preserving Data Mining (PPDM) performed at data publishing is commonly termed Privacy Preserving Data Publishing (PPDP) [16].

The initial rationale of performing PPDM was the extension of the traditional data mining techniques to modify data to mask sensitive information. PPDM solutions were closely linked with existing data mining algorithms that are available. With PPDP, the objective is how to publish data that is still useful for data mining purposes, whereas PPDM involves the actual data mining task [17].

2.3.3 Data De-identification

To realise the benefits of sharing data whilst still maintaining privacy, de-identification can be used. De-identification is a technique that removes obvious identifying information from disclosed records [18]. One approach to aid in the legitimate and authorised disclosure of personal health information is to de-identify data prior to it being used or at the earliest opportunity [19].

De-identification occurs as a result of removing personal information which is used (whether directly or indirectly) to identify individuals in a data set. De-identification is a technique used to remove personal information of which pseudonymisation and anonymisation are a type of de-identification method. These techniques are used to reduce the risk of identifying individuals from an unsecured data set. Within the expanse of data privacy, these terms are often misrepresented and interchanged in their use with no consensus of the terminology used [20].

Currently, the most frequently used application of PPDM techniques relates to the release of health care data related to patient information. In the United States (US) this legislation is referred to as the Health Insurance Portability and Accountability Act (HIPAA) [4]. The HIPAA is discussed in further detail in the next section.

2.3.4 Data Anonymisation and Data Pseudonymisation

Data anonymisation can be defined as a method of sanitising information with the aim of protecting privacy whilst at the same time retaining the ease of analysis.

Pseudonymisation is when the aim is to reverse the process of anonymisation and a variable or identifier is used to replace the original data [21]. As this study suggests, when using pseudonymisation, it can be difficult to keep the usefulness of the information when processing it whilst at the same time ensure anonymisation of the data to a sufficient level of privacy. By “replacing an attribute with another”, which occurs with pseudonymisation, it allows the logical integrity of the source of the individual data to be kept by being able to identify a particular individual if necessary.

However, the individual is not able to be directly identified as the data is obfuscated or masked.

This study will focus on commonly used and data anonymisation techniques that have been proposed in studies involving the anonymisation of health care information. The full anonymisation process has many techniques that are available, namely generalisation, suppression, micro-aggregation, and subsampling. Generalisation, as quoted by Sweeney et al. [11], “involves replacing (or recoding) a value with a less specific but semantically consistent value. Suppression involves not releasing a value at all.” Sweeney is one of the forerunners in research pertaining to data anonymisation. It was further mentioned by [11] that a combination of generalisation and suppression is a common method used by the k-anonymity privacy model. Micro-aggregation and subsampling are not part of the scope of this research as a technique of choice.

In contrast, the converse of anonymisation is called re-identification, or sometimes referred to as de-anonymisation. The next section discusses this further.

2.3.5 Re-identification

The reverse process of de-identification is re-identification, which happens when a single record within a data set can be identified. The simplest example of a form of re-identification is called identity disclosure. Identity disclosure occurs when a match is made in a de-identified data set with a record in a publicly known data set [18]. Further detail on the various methods of privacy threats are outlined in Chapter 3.

Furthermore, various data utility metrics have been proposed to measure risk. These metrics are very important so that data utility loss can be quantified and as a result be used to assess the level of de-identification in a created dataset [20].

To this end, by using metrics that have been proposed in other research areas involving data anonymisation, it is possible to use the same methods to establish the likelihood of re-identifying a person’s information based on a specific data set. This

metric of re-identification is important in this study as it will be used in testing the hypothesis as defined.

2.4 International Legislation Surrounding Data Privacy

Data protection is a hot topic amongst developed nations around the world. Each country has various levels of maturity when it comes to data protection, and this is mainly as a result of the various timelines for when data protection laws were introduced and enforced.

This section outlines the more popular data protection legislation and laws that are currently in force around the world. These are the most comprehensive set of laws relating to data protection that is currently available.

2.4.1 Protection of Personal Information (POPI) Act: South Africa

The Protection of Personal Information Act (POPI) Act was promulgated (or signed into law) in November 2013 [2]. This is South Africa's first data protection legislation. It is, however, currently not 'effective' as yet. As soon as the POPI Act is effective, organisations and institutions in South Africa will be afforded a one-year grace period to comply with POPIA. Regulations relating to the POPI Act were tabled in parliament on 3 December 2018 and were published in the Government Gazette on 14 December 2018 [22].

As mentioned in the previous section, the POPI act specifically contains principles that govern how data should be stored, used and processed, and more specifically relates to the security of the sensitive data [2] [23] . The principles are:

1. "Accountability: The responsible party must ensure that the principles are adhered to;
2. Processing Limitation: There must be limits to the processing of information; processing must be lawful and not excessive;
3. Purpose Specification: Personal information must be collected for a specific, defined and lawful purpose that is related to the responsible party's activity; the subject should be aware of this purpose;

4. Further Processing Limitation: Any further processing must be compatible with the purpose that the information was collected for;
5. Information Quality: The responsible party must ensure that the personal information is complete, accurate and not misleading; the information can be updated if necessary;
6. Openness: A notification must be given to the Information Protection Regulator before the information is processed the subject must be notified that data is being collected about them;
7. Security Safeguards: The responsible party must ensure that the integrity of the collected personal information is maintained;
8. Data Subject Participation: The subject has the right to ask and be given the details of any information on him/her that the responsible party might have, at no cost."

There are severe consequences for noncompliance of the POPI Act [2][24]. When conducting business within South Africa, the POPI Act will dictate the steps on how organisations would go about processing personal information that is compliant with the Act. The principles of the POPI Act outlined these considerations.

Furthermore, to ensure enforcement and promote the rights protected by POPIA, the Act states that the government must appoint an Information Regulator. An information regulator was appointed in November 2016 to fulfil this purpose as regulation without enforcement is not efficient [25].

Should an organisation be investigated by the Information Regulator and found in violation of the POPI Act, the following four consequences could result [2][26] :

1. Administrative fines levied on organisations of up to ten million rand;
2. Criminal prosecution with fines of up to ten million rand and the possible inclusion of a prison term of up to twelve months;
3. Issue of an "enforcement notice" to immediately cease the processing personal information; and
4. Initiation of a civil action lawsuit representing individuals or groups.

POPIA, however, also has the following exclusions when referring to violations of the Act [24]:

- When information is processed in a personal capacity and not in a commercial environment;
- De-identified data where the application of anonymisation techniques were used correctly; and
- When information collected enhances the safety of the public at large or is in the interest of national security [2].

2.4.2 African Privacy Laws

There are different data protection laws in Africa and is increasing as a result of more and more nations conducting their business on a global scale and expanding. As of 2017, there are an estimated sixteen African countries that have enacted data protection legislation, five countries have initiated have data protection bills and nine have intentions to enact data protection bills. However this process has not been fully completed in these countries, as in the case of South Africa [23].

A comparison of selected African privacy laws compared to the POPI Act is shown below, taken from [23]:

Country	Act	PoPI Principles								Other Areas					
		Accountability	Processing Limitation	Purpose Specification	Further Processing Limitation	Information Quality	Openness	Security Safeguards	Data Subject Participation	DPO Required	Breach Notification	Cross-border Data Transfer Limitations	Electronic Marketing	Online Privacy	Enacted Year
South Africa	PoPI	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		2013
Angola	PDL		✓	✓		✓		✓	✓			✓		✓	2011
Benin	PPD		✓	✓	✓	✓	✓	✓	✓		✓	✓	✓		2009
Burkina Faso	PPD		✓	✓		✓		✓	✓			✓			2004
Cape Verde	PPD		✓	✓		✓		✓	✓			✓	✓	✓	2013
Gabon	PPD		✓	✓		✓		✓	✓			✓			2011
Ghana	DPA		✓	✓	✓	✓	✓	✓	✓		✓				2012
Ivory Coast	PPD		✓	✓		✓		✓	✓	✓		✓			2013
Madagascar	PPD		✓	✓		✓		✓	✓	✓		✓			2015
Mali	PPD		✓	✓		✓		✓	✓			✓			2013
Mauritius	DPA		✓	✓	✓	✓	✓	✓	✓			✓			2004
Morocco	PIRP PD		✓	✓		✓		✓	✓			✓	✓		2009
Senegal	PPD		✓	✓		✓		✓	✓			✓			2008
Seychelles	DPA		✓	✓	✓	✓	✓	✓	✓			✓			2003
Tunisia	DPA		✓	✓	✓	✓	✓	✓	✓	✓		✓			2004

Figure 2: Comparison of selected African privacy laws with POPIA [23]

The relevance for including the comparison of other African privacy laws is often that data and personal information is moved across borders to other regions by organisations and businesses, and as a result one should be familiar with privacy laws of these regions. Doing business within Africa is no exception. It is always necessary to have this information known so that adequate privacy protection levels can be put in place by South African businesses when conducting business in Africa [23].

2.4.3 General Data Protection Regulation (GDPR): European Union

On 25 May 2018, the General Data Protection Regulation (also called the GDPR) became law. The introduction of GDPR was done to “modernise laws that protect the personal information of individuals” [27]. The European Union (EU) recently adopted the GDPR which is considered one of the strictest data protection rules [28].

Previous data protection laws in the EU were in force since the 1990's and had become very outdated [29]. Changes to the world we live in with regards to people, processes, and technology have necessitated the changes to privacy laws so that current needs are met. The GDPR was a direct replacement of the data protection directive of 1995, which makes it the new framework in Europe for data protection [30]. The GDPR standardises laws across European countries and aims to strengthen the data rights of individuals[31]. The Information Commissioner's Office has been tasked with enforcing the law in the UK [24].

The GDPR supersedes all existing national laws that are currently in force by EU members [30]. It includes changes to the previous Data Protection law (Directive 95/46/CE) that includes high penalties if compliance is not enforced within companies [30]. Organisations need to be internally structured so that there is appropriate accountability placed on individuals within the organisations by having a data protection officer (DPO), and also requiring immediate notification of data breaches to regulators within certain timeframes[27]. With GDPR, by needing to obtain an individual's consent to process their personal information has also made the processing of personal information more difficult than previously [30].

With the transitional period of two years that was given to organisations in the EU to become compliant between 2016 and 2018, this allowed for the interrogation of the new GDPR principles by companies to make room for an amendment of company policies and procedures in order to be in line with the new GDPR Act.

2.4.4 Health Insurance Portability and Accountability Act (HIPAA): The United States of America

In the United States (US), there are multiple patchwork laws that make up regulations for the processing of personal information [32]. As there is an overlap of federal and state laws, these laws do not provide thorough and extensive protection over data rules relating to individuals as is required.

Below are the two most commonly known US federal privacy laws [32]:

- The Federal Trade Commission Act (FTC Act) is a federal law that protects consumers. This law is intended to stop the unfair practices associated with offline and online security policies relating to privacy.
- The Health Insurance Portability and Accountability Act (HIPAA) released in 1996 regulates the release of medical data. The HIPAA Privacy Rule specifically applies to “the collection and use of protected health information (PHI)” and contains standards that must be considered when protecting medical data.

The HIPAA requires an institution to notify data subjects of their rights to privacy and the institutions’ requisite privacy practices [32]. The HIPAA Privacy Rule went a step further to specify 18 fields of data for generalisation or removal from the data set [3].

An important update was made to the HIPAA in September 2013, with the implementation of the ’Omnibus Rule’ [33] . The aim of the ‘Omnibus Rule’ was to tighten and strengthen the personal privacy of individuals. It also expanded the definition of “covered businesses”. A covered business with regards to the Omnibus Rule references all companies that create, collect, store, or transmit PHI on behalf of another covered entity. This, therefore, provides all-round protection of personal information.

2.5 Data Breaches

When data is released to external parties outside an organisation, measures must be taken to ensure that the data is in an anonymised form. However, there are occasions where even though anonymisation was applied, it was not sufficient. One of the most common attacks when it comes to data breaches are called inference attacks where a combination of attributes could be used to identify an individual. As mentioned in [34], Sweeney successfully showed that by using a combination of address code, respondents’ gender, and respondents’ birthdate, 87% of the United States of America’s (US) population could be identified [34].

In this section, a few known data security breaches are outlined.

2.5.1 The Netflix Prize

Netflix is a service that lets users view TV shows, movies, and documentaries that are streamed via the internet into their homes [35].

In 2006, Netflix ran a competition and released almost half a million records publicly, which contained the movie-rating preferences of anonymised users. There were 100 million movie ratings released purposely as a source of information for the competition called The Netflix Prize. Information that could identify customers was removed and contestants had to predict which movies customers would prefer. To see if the predictions were correct, they were compared to how customers rated those movies in real life [36].

After the release of this data, the data set was de-anonymised successfully [36]. The researchers in this article could identify some on the 'anonymised' users with ease, if some knowledge about information pertaining to individuals within the data set was known. Prior to the release, Netflix had consulted with computer scientists in the field to anonymise the data. In this scenario the appropriate action to ensure the privacy of the data set was taken, however, it was not adequate. Common personal information included in the Netflix release was the same as the study by Sweeney where an individual's gender, address code, and birthdate could identify up to 87% of the population of the US [34].

2.5.2 The Personal Genome Project

The Personal Genome Project (PGP) was started in 2006. One of the aims of the project was "to sequence the genotypic and phenotypic information of volunteers and display it publicly online in an extensive public database." The volunteers were informed prior to the research being done in order to obtain their consent [37].

This project resulted in another breach of personal information in 2013 that was identified by Sweeney, together with other researchers. This study was able to link an individual's name and their contact details to publicly available profiles of individuals

that took part in the Personal Genome Project. The publicly available profiles from the Personal Genome Project contained private medical information, genomic information, as well as demographic information, such as birthdate, sex, and address codes. Demographic information found was linked to public records such as voter lists, thereby resulting in identifying 84% to 97% of the profiles for which names were provided [38].

2.6 Related Work

A common use for data anonymisation is within the area of data publishing of sensitive information by governmental organisations around the world. The Public Use Microdata Sample (PUMS) files provided by U.S. Census Bureau is an example of such publishing. The data set was de-identified prior to release for statistical purposes. Other statistical agencies also perform a similar function with their data, including the Confidentialised Unit Record Files (CURF) provided by the Australian Bureau of Statistics (ABS) and Statistics New Zealand [39].

There are however a few limitations to consider when anonymising raw health data in preparation for data publishing. When using the k-anonymity and l-diversity privacy models, it is difficult to specify the quasi-identifiers upfront when it is unknown what information adversaries that would attack the data already have. Furthermore, data mining results that is done on already anonymised data could be potentially different from the original data [39]. In the examples mentioned above, it is rarely possible to use the output data set for meaningful data mining as it will differ vastly from the original dataset. To this end, there are many studies that considers the topic of protecting sensitive data whilst still maintaining data utility. Some of the work in this area includes, among others [1], [40] and [7]. In this study, it is proposed to measure and identify those privacy models that can be used to gain the most utility of a data set within the South African context of POPIA.

Personal health information is another popular area where information about a person is most sensitive and contains intimate details, be it physical health or mental health. Therefore, confidentiality of this data must be protected at all costs and exposure of this trusted information must be avoided. Conversely, by de-identifying personal health

information, it can be used for secondary purposes that are important, like research in the health field [41]. As mentioned in the previous section, the HIPAA Act of 1996 legislation specifically defined methods for de-identifying health data [5].

Whilst the articles above contribute to advancing the concepts of data anonymisation, generalisation and suppression, it unfortunately does not address the practical application of these privacy models and concepts. In Spengler et al [42], an important aim of the article was to bridge the gap between legislation pertaining to data privacy and the practical, technical solutions that are required. In that article, the open source data anonymisation tool ARX was used to implement anonymisation by using generalisation and k-anonymity. Health related data was used for the purposes of that study [42].

Alternatively in [43], the technique of k-anonymity was proposed to create k-anonymous tables from a set of customer records. This is done in a distributed scenario where there is a miner that wants to mine the entire table of customer records. Whilst the articles major contribution was a use case for the anonymisation of customer data, the author does not consider the anonymisation of customer data in a practical way as this study suggests. With the inception of POPIA and taking that into consideration, data anonymisation scenarios in a South African context is also fairly new.

2.7 Summary

This chapter began with a discussion of the data privacy initiative and the motivation and rationale for performing data anonymisation. A brief overview of the data anonymisation and de-identification concepts was then outlined in order to provide a background to privacy preserving data mining and privacy preserving data publishing. Legislation and laws surrounding data protection was discussed with examples of data breaches that have previously occurred. Data privacy regarding data utility and attacker models was outlined. Finally, other related work in this area of data anonymisation were mentioned. In the next chapter, privacy models, techniques, threats, and their associated tools are discussed in detail.

3 DATA ANONYMISATION: PRIVACY MODELS, TECHNIQUES, TOOLS AND PRIVACY THREATS

In the previous chapter, it was shown that there exists a need for maintaining data privacy by focusing on data anonymisation. Legislation and laws have been introduced to ensure that the privacy of individuals is maintained.

In this chapter, the theoretical understanding of the various privacy models, techniques, and tools that are available are examined. Firstly, the data anonymisation techniques are presented showing the two broad categories that are available when performing anonymisation. Thereafter, a description of common privacy threats with examples are provided. Popular privacy models are explained in order to show the various ways that these anonymisation techniques can be implemented within privacy models. Furthermore, important parameters for resolving data privacy issues are discussed, namely data utility and attacker models. Finally, a summary of the suite of anonymisation tools and technologies that are available to perform anonymisation are presented.

3.1 Data Anonymisation Techniques

Privacy Preserving Data Mining (PPDM) techniques can be implemented in a way to ensure privacy but also extract knowledge contained in the data set [14]. The utility of the data is maximised because the guarantee of privacy levels is built into these techniques. As a result, data mining actions can still be sufficiently performed on the transformed data.

Anonymisation of a data set before the data set is published can also be done. PPDM that occurs at the same time or prior to data publishing is referred to as Privacy Preserving Data Publishing (PPDP) [14]. This study will focus on anonymising a data set prior to it being used by a third party or external organisation; therefore, the focus will be on PPDP [39].

To achieve this goal of eliminating the ability to identify an individual from personal data sets, data anonymisation techniques are needed to accomplish these tasks. Data anonymisation techniques fall into two broad categories:

1. Perturbative
2. Non-Perturbative

These two categories are described in further detail.

3.1.1 Perturbative Techniques

Perturbation involves replacing the original values in a data set for values with the same statistical information [14]. As a result, this method extorts the original data. There are many types of perturbative techniques, with the most common being the addition of noise into a data set as well as data swapping. Both techniques are described below.

3.1.1.1 Adding Noise to a Data set

The use of perturbation necessitates introducing an external factor such as “noise" into the data in order to mask individual values [44]. One method to achieve this is by using a technique called randomisation, where noise that is added is so large that it is difficult to recognise the individual records in a data set. A disadvantage to this is that perturbative techniques distort the data and do not preserve the truthfulness of the data set thereby making it nearly impossible to obtain accurate statistical information as a result [44], [45].

3.1.1.2 Swapping values in a data set

Data swapping involves swapping values across a data set to maintain the preservation of privacy [44], [46]. As some databases are used for statistical purposes, the aggregation (or aggregate characteristics like averages and totals) of the data is important [45]. Therefore, the use of swapping values is a better option for statistical purposes. These data transformations, when the swapping is complete, map back to the original data and exhibits the similar properties. As data swapping is a data

transformation technique, the resultant data can be used for statistical tabulation as all data is released [46]. An important advantage is that data swapping does not have the same weakness as randomisation in that it guarantees that statistics will be preserved.

3.1.2 Non-Perturbative Techniques

Non-perturbation techniques involve reducing the granularity of the individual data set to achieve privacy [47]. In this way the original data set is kept the same and is not altered, therefore, no extra “noise” is added as with perturbative techniques. An important outcome of not introducing additional “noise” into a data set is that the truthfulness of the data can be preserved. Two techniques that are commonly used because they do not disturb the original data set is the use of a generalisation and a suppression mechanism. Both are described below.

3.1.2.1 Generalisation:

Generalisation involves replacing a value within a data set with a less specific value that is semantically consistent [11]. This is sometimes known as recoding. A few terms must first be explained to understand how the concept works.

Firstly, in a data set, an attribute represents each column in the data set and contains all possible set of values in that domain [11]. Attributes in a table are unique to each other. Generalisation is commonly performed using a generalisation hierarchy to transform the attributes. In a generalisation hierarchy, the uppermost parent node is the highest general value within that domain. As the hierarchy is split into its respective leaf nodes downwards, the values become more granular and specific. In this way values higher up in the hierarchy are grouped into larger numbers to distort the ability of an attacker to gain knowledge of a specific value.

Secondly, a direct identifier is an attribute that is highly distinguishable and can be used to explicitly identify a record [11]. Some of these attributes include name, address, and identity number, just to name a few. As these attributes are too obvious

in identifying a record, it is usually suppressed from a data set to limit the release of private information.

Thirdly, the term quasi-identifier (QID) refers to those attributes that together may be linked to identify an individual [12]. Quasi-identifiers are commonly referred to as indirect identifiers. On their own these attributes cannot be used as an identifier, however, when used in combination it is possible to use them for linkage. Moreover, very often these attributes are needed in the data set as they are useful in the analysis that will be performed on the data set. Therefore, during the data anonymisation process, quasi-identifiers are transformed thereby ensuring that privacy needs are met.

Finally, the last term to understand when performing generalisation is called sensitive attributes or SA [48]. Sensitive attributes are those attributes that are personal and that most individuals do not want shared, for example disease diagnosis, salary, disability status and the like.

Taken from [12], below is an example of how a generalisation hierarchy is derived.

Table 2: Data set with Quasi-identifiers Before and After Transformation using Generalisation [12]

Before Generalisation		**After Generalisation**	
Age	**Gender**	**Age**	**Gender**
34	Male	20 - 39	Male
22	Female	20 - 39	Female
66	Male	60 - 79	Male
70	Male	60 - 79	Male
35	Female	20 - 39	Female
21	Male	20 - 39	Male
18	Female	*	Female
19	Female	*	Female

In Table 2 above, the example data set contains two attributes, age and gender. Both attributes are classified as quasi-identifiers. After the generalisation, the quasi-identifier Age has been transformed into age groups of twenty per group.

The generalisation hierarchy for the attribute Age is shown in Figure 3 below. The attribute has been grouped with a parent node at the highest point, and then gradually reduced into smaller, more specific groupings in each of the leaf nodes. The quasi-identifier Gender could only be suppressed.

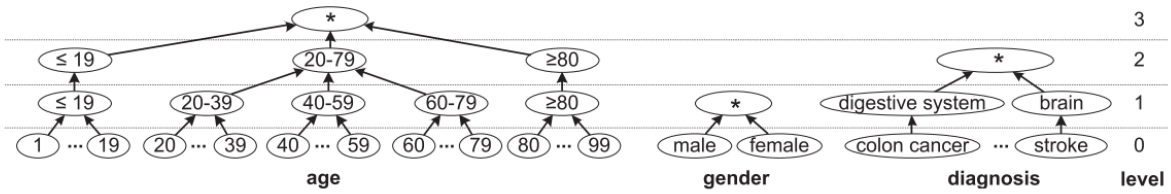


Figure 3: Generalisation Hierarchy for quasi-identifier Age [12]

A key disadvantage to using generalisation is that the generalisation hierarchy must be manually setup and generated for all the quasi-identifiers present [49].

3.1.2.2 Suppression:

When suppression is used, a value is not released at all [11]. It suppresses the entire value of that attribute and that value is, therefore, removed from the data set [50]. The resultant value becomes a null value.

In reference to the Table 2 and Figure 3 above, the attribute Gender is transformed using suppression. Suppression is commonly used when the release of a specific value can lead to a significant breach of the privacy of that individual record. The different types of suppression referred to in [50], [51] are:

- attribute suppression – when the whole values are suppressed
- record suppression – when the full row or record is suppressed
- value suppression – when a certain value in a data set is suppressed
- cell suppression – suppresses some records in a data set

- multidimensional suppression – when values in certain records are suppressed, in relation to other values in the data set.

In summary, there are privacy models that make use of generalisation and suppression to enforce privacy. Further details on these privacy models are outlined in the next section.

3.2 Privacy Threats

When anonymising structured data, a general method of attack assumes the linkage of a protected data set with a public data set (or similar background knowledge about individuals that are known) [52].

There are three common types of privacy threats, namely [53]:

- Membership Disclosure
- Attribute Disclosure
- Identity Disclosure

A brief summary of each privacy threat is outlined below.

3.2.1 Membership Disclosure

With membership disclosure, linkage of the data could allow an attacker to learn whether one's information is included in a data set. This becomes a problem when the selected data set contains information of sensitive attributes that are the same, for example only diagnoses of diabetes patients or cancer patients. This sensitive information can be inferred and revealed by an attacker [53].

3.2.2 Attribute Disclosure

Attribute disclosure occurs when new information about a data subject that is contained in a data set is revealed to an unauthorised party. By viewing the data that has been released, an attacker would be able to infer the attributes of an individual quite easily than would have been possible prior to the data release [53]. In other

words, sensitive attributes that an individual would not want disclosed must be protected from an attacker [54].

3.2.3 Identity Disclosure

With identity disclosure, a link can be made between an individual and a specific record in a data set quite easily. This is also referred to as re-identification. With this type of attack the institutions that are supposed to protect the data in question could face serious consequences as a result of this data leak, which is a result of privacy regulations that are not met or enforced [52].

A simple visual example of these three types of privacy threats taken from [52] are shown in Figure 4.

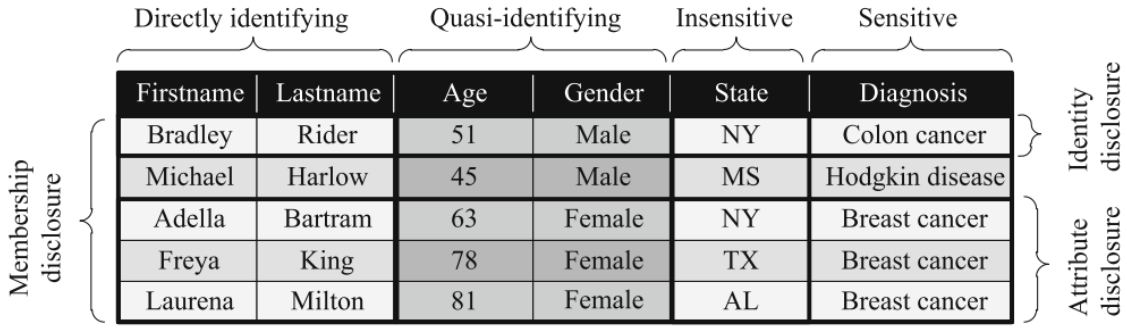


Figure 4: Types of attributes with types of disclosure [52]

3.3 Privacy Models

The transformation of data inevitably leads to a loss of information [55]. In order to avoid this, a balance must be sought between an increase in privacy protection on one side and a decrease in data quality on the other.

When disclosing personal information in data sets, there is always the risk of re-identification. Privacy models ensure that privacy requirements are met, whilst the implementation of the privacy model is performed using data transformation, thereby preserving utility [48]. This section provides further information on the privacy models that are available.

3.3.1 K-anonymity

Finding and accessing information in this digital age has become very easy. Therefore, the privacy considerations of individual subjects have taken on an importance not considered previously. A common practice where this is used is when large databases are used to store personal information of a sensitive nature but identifiers are removed to ensure privacy [12].

Consequently, this released information can be linked with the individual's corresponding known information, possibly from other databases that are publicly available. As such, the privacy of an individual's information can potentially be violated. As mentioned previously, Sweeney [56] demonstrated it is possible to find out who has a certain disease using a publicly available health database linked together with voter lists. As a possible solution for the problem, Samarati and Sweeney [57] were the first to propose a technique called k- anonymisation.

The k-anonymity principle was first used in protecting the association of a patient's record in publicly available data directly to a patient. Data from these two records can be interpreted and then triangulated to obtain the personal information of an individual, thereby disclosing their identity [58].

K-anonymity requires that "each record in a data set is indistinguishable from at least k − 1 other records regarding attributes which could be used for re-identification attacks" [59]. In the following scenario, if k =3 and the potentially identifying variables are date of birth and gender, the resultant k-anonymised data set will have at least three records for each combination of date of birth and gender. As a result, this type of anonymisation reduces any risks of re-identification to data subjects, for example, where there is a data breach and the data cannot be linked to any specific individual [60].

According to [12], data is transformed via two methods when using k-anonymity: (a) firstly, the generalisation of attribute values, and thereafter (b), the data records are suppressed. Generalisation results in data that is well-suited for analyses by scientists

who study diseases within populations, while suppression significantly reduces loss of information. [12] also found that the combination of generalisation and suppression results in a significant increase in data utility. Consequently, in this study, generalisation and the suppression of data will be used to determine whether it is an effective approach to transform the data in question.

In the k-anonymisation process, it is important to identify key attributes. As previously mentioned, direct identifiers like first name, social security number, and identity numbers must be removed from personal records. These are explicitly identifying in nature. There are other types of attributes known as quasi-identifiers (also known as pseudo-identifiers) that can be contained in a given data set, which could possibly be used to infer the identity of an individual. It can be seen that by combining these various attributes in a data set, the possibility of narrowing down the options to a smaller group of individuals is possible. A popular example of quasi-identifiers are age, zip-code, and gender that are available publicly in census records [44].

There are two important definitions for k-anonymity that must be reviewed.

The first definition as mentioned in [11]:
"QI being a quasi-identifier for a given table U with
T(A1 . . . An), fC : U → T, fg : T →U', where U ⊆ U', a quasi-identifier of T(QT) is a set of attributes {Ai . . . AJ}⊆ {A1 . . . An}, where ∃pi ∈ U such that fg (fc(pi)[QT]) = pi"

The second definition as stated in [54] :
"a table T satisfies k-anonymity if for every tuple t ∈ T there exist k − 1 other tuples ti1ti2 . . . tk−1 ∈ T such that ti1[C] = ti2[C] = . . . tik−1[C] for all C ∈ QT"

From the definitions above, k-anonymity can, in simple terms, be defined as:
"Each release of the data must be such that every combination of values of quasi-identifiers can be indistinguishably matched to at least k respondent [44]."

Figure 5 below graphically represents the famous example used by Sweeney [56] to relink a data subject using quasi-identifiers from two data sets. The data subject in question was the then Governor of Massachusetts from the United States of America.

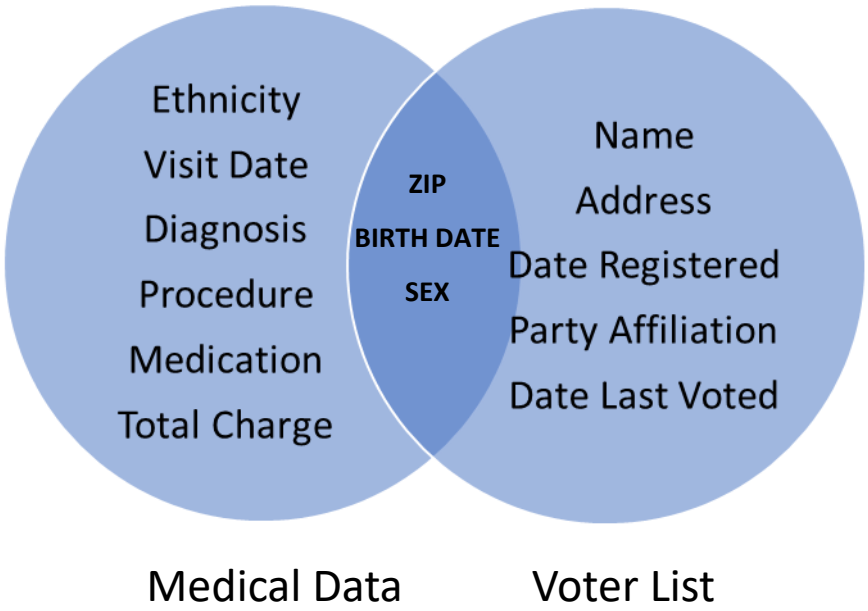


Figure 5: Example of relinking individuals from two separate data sets [56]

As was found in the study by Sweeney [56] in 1997, William Weld was the governor of Massachusetts when his medical records were contained in the Group Insurance Commission(GIC) data set that was released. The GIC collected health insurance for government employees in Massachusetts and sold a copy of this health information which was perceived to be anonymous [61]. Governor Weld had publicly confirmed that identifiers were deleted before release and coincidentally he was hospitalised prior to the data release in Cambridge, Massachusetts. Sweeney then purchased the voter list which contained names, addresses, and similar information to the GIC data set, namely zip code, birthdate, and sex (or gender). When a correlation was done, only six people had the same date of birth as Governor Weld with only three of them being males. It was then narrowed down to a single record as he was the only one in his zip code (or address code). As a result, the Governor's personal information was easily identifiable.

The privacy breach above in which medical data that was a part of insurance data that already had direct identifiers removed prior to release assisted in identifying Governor Weld has had an important impact on the exploration and development of privacy laws such as the HIPAA [61]. In the above example, this form of privacy breach is called identity disclosure.

Although k-anonymity is a popular method to use to perform data anonymisation due to the conceptual simplicity in the application of the algorithm as well as there being many algorithms available to perform the anonymisation [44], this privacy model is vulnerable to many other types of breaches especially when the attacker has some form of background knowledge. Other examples for these types of attacks are taken from [44], [54] and are briefly mentioned below:

- Homogeneity attacks occur when a sensitive attribute contains the same value for all the records for that attribute; therefore, it is easily distinguishable even after it is k-anonymised;
- Background knowledge attacks occur when a link between quasi-identifiers and a sensitive attribute enables an attacker to narrow the options of values that are available for the sensitive attribute, thereby learning information about an individual record.

3.3.2 L-diversity

The l-diversity privacy model was developed to cater for the shortcomings in the k-anonymity model [44]. One reason for developing a newer model was that protecting the disclosure to the level of k-individuals did not protect the privacy of corresponding sensitive values sufficiently, especially so in circumstances where the sensitive values were similar [44]. A sensitive attribute (SA) can be described as an attribute for an individual in a data set that must be kept confidential and private from people who do not have access to or are not allowed to view the original data set.

The first definition of l-diversity as in [44] states:
"Let *a $q*$-block be a set of tuples such that its non-sensitive values generalize to $q*$. A $q*$-block is l-diverse if it contains l 'well represented' values for the sensitive attribute S. A table is l-diverse, if every $q*$- block in it is l-diverse."*

A simpler definition of l-diversity as describe by Machanavajjhala et al [54]:
"An equivalence class is said to have l-diversity if there are at least l 'well-represented' values for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity."

Consequently, the focus of l-diversity is to preserve the diversity amongst the sensitive attributes as well as maintain the minimum group size. In this way, l-diversity can provide a level of privacy for data when the data publisher is not aware of the knowledge that an external adversary might know [54]. L-diversity provides privacy by ensuring that values of sensitive attributes are well balanced and dispersed in such a way that privacy breaches can be prevented.

The l-diversity privacy model also has limitations. The first limitation pertaining to l-diversity is its inadequacy in the way it assumes adversarial knowledge. As the research by [62] indicated, if an adversary has knowledge of the global distribution of a sensitive attribute, it is possible to find out specific information on an individual within that group.

The second limitation of the l-diversity privacy model is its inadequacy in preventing attribute disclosure. Below are two types of attacks that can occur when using the l-diversity privacy model [63]:

- Skewness Attack**:** occurs when the distribution of a data set is skewed as a result of applying the l-diversity privacy model. There l-diversity is not enough to prevent attribute disclosure.

- Similarity Attack: occurs when the sensitive attributes are very similar although they are distinct. Important knowledge can be gained from the similarities found within the equivalence class.

Due to the fact that l-diversity aims to diversify the sensitive values in a data set, sensitive information can unintentionally be leaked due to the semantic closeness of the sensitive values in the group [62].

3.3.3 T-closeness

A newer privacy model, the t-closeness algorithm, was developed to address vulnerabilities that were identified in previously mentioned privacy models.

According to [64], the application of k -anonymity and l-diversity is often not adequate for the protection of numerical attributes. As in the example of the attribute for a salary of an individual, even if l -diversity is fulfilled, sensitive information can still be revealed if the salary values fall within a narrow range which still, therefore, leaks personal information.

According to [62], in the t-closeness privacy model, the distribution of a sensitive attribute must be similar to that same attribute's distribution in the entire data set. By doing this, specific sensitive information can still be protected.
Therefore, the distance between the two distributions should be no more than a threshold (t). [62] also proposed using the Earth Mover Distance measure for t-closeness. In order to introduce and manage gaps between values of sensitive attributes, t-closeness uses the Earth Mover Distance metric [64]. It receives the precise distance between two distributions and takes into consideration the semantic closeness. This method allows the data collector to use other anonymisation techniques besides generalisation and suppression [65].

3.3.4 Privacy Model Selection

It can clearly be seen from the above that each privacy model has advantages as well as disadvantages. A choice of a privacy model for a particular purpose differs from scenario to scenario. As a result, there is no clear method to determine the appropriate data anonymisation privacy model to be used for this study.

3.4 Generalisation Hierarchies

Generalisation hierarchies are typically used in de-identifying data sets with personally identifiable information. During the anonymisation of data quasi-identifier attributes are transformed in such a way that privacy needs are met. This data transformation (called recoding) is mainly done with the use of generalisation hierarchies.

As seen in Table 2 previously, when generalising values of the attribute Age, the values are transformed into groups by age and thereafter suppressed, whereas the attribute Gender was only suppressed. The result was a generalised hierarchy as shown in Figure 3. Generalisation hierarchies are used primarily for categorical attributes [12], but can also be used for continuous attributes by performing categorisation.

Defining bigger intervals (fewer levels of hierarchies) will decrease the risk of re-identification; however, if this attribute is relevant for an analysis, defining bigger intervals will result in the data set losing utility [12]. In brief, care must be taken when generalisation hierarchies are first created.

Due to the importance of generalisation hierarchies when anonymising data using k-anonymity, in this study the focus is on attribute generalisation by using user-specified hierarchies. User-specified hierarchies are used to describe the rules for replacing values with more general but semantically consistent values on increasing levels of generalisation.

3.5 Anonymisation Tools

This section presents an overview of common anonymisation tools that are available on the market. These tools are either open-source or commercially available products that need to be purchased and require a licence. A comprehensive overview of each of these anonymisation tools are presented, together with highlights of the functionality in each.

3.5.1 ARX Anonymisation Tool

The ARX anonymisation tool [66] is an open source software tool that can be used to anonymise personal information into a form that can be shared with the appropriate level of anonymisation already applied. ARX has been primarily created for the use of de-identifying biomedical data and is the most comprehensive tool for applying a variety of anonymisation methods [55].

ARX is highly configurable and a few of the important characteristics of ARX are highlighted below:

- ARX can implement a variety of privacy models, some of which are k-anonymity, t-closeness, l-diversity, and l-presence;
- A key characteristic of ARX is that it minimises disclosure risk whilst also reducing information loss on the data output;
- Utility measures can compare the outputs of various data transformations that are performed. Information loss as a measure is directly an opposite measure of utility [12];
- ARX has a user-friendly intuitive graphical interface;
- ARX is capable of importing data from various sources, for example comma separate value (CSV) files, Microsoft Excel files as well as Microsoft SQL Server, just to name a few.

Within ARX, data is transformed using generalisation hierarchies [58]. This is done by using a wizard to build the required generalisation hierarchies. The process of creating, modifying and managing hierarchies are easy to do in ARX. Privacy models and coding models are also easy to adjust according to the user needs.

The use of a graphical interface to show the statistical results is very useful to users who are new to data anonymisation. A side-by-side view of the original data set next to the anonymised data set shows a visual representation of the differences to the data post-anonymisation. Also, after the anonymisation is complete, the risk analysis can easily be done with the built-in view. An important aspect is that the view shows

details relating to the estimated re-identification risks as a result of running anonymisation using the different privacy models.

3.5.2 Other Anonymisation Tools

Tools for de-identifying data are available on the market according to specified needs. As described in [67], there are five generally available de-identification tools in the market, but this list is constantly expanding as it is an ever-growing field:

- The **PARAT** tool from Privacy Analytics Inc.
- The **µ-Argus** tool, from the Netherlands National Statistical Agency.
- The Cornell Anonymisation Toolkit (**CAT**), also called Incognito implemented by Cornell University
- The University of Texas (Dallas) (**UTD**) Anonymisation Toolbox
- The **sdcMicro** package in R

This following section provides a brief overview of each of the tools mentioned above.

3.5.3 Privacy Analytics Risk Assessment Tool (PARAT)

Privacy Analytics Risk Assessment Tool (PARAT) from Privacy Analytics Inc. is one of the most popular de-identification software tools on the market. Only limited details are available publicly as it is closed-source [5]. PARAT has been used by seven of the top ten Fortune 500 healthcare companies [68]. The PARAT tool offers protection and privacy for three types of identity disclosure risks.

PARAT supports a Windows platform and is compatible with various database types such as Microsoft SQL databases and Oracle databases. There are multiple steps for the anonymisation process in PARAT. Firstly, indirect identifiers are specified in a data set. Thereafter, a re-identification risk threshold is selected. PARAT then does a risk analysis on the proposed re-identification risk using three attacker models, namely

prosecutor attacker model, journalist attacker model and marketer attacker model. Finally, the re-identification risk is reduced to an acceptable level by applying various de-identification techniques[69]. A recent version of PARAT version 6.0 was released in 2014 [68].

3.5.4 μ -Argus

μ -Argus is a closed-source software tool that implements various popular techniques [67]. It was developed by Statistic Netherlands. The name “Argus” is an acronym for "Anti-Re-identification General Utility System" [69]. The software supports Windows and Linux Ubuntu and was developed by the European Union within the Computational Aspects of Statistical Confidentiality project. The most recent version 5.1.3 was released in March 2018 [70].

According to [69], the first step in the process is to identify indirect identifiers in a data set. The software then determines the re-identification risk for each record in the data set. Rarity of the population within the data set is estimated and the risk of re-identification based on available combinations of variables is calculated. Lastly, unsafe combinations of variables are identified, and manual global recoding done. As soon as the number of unsafe combinations has been reduced, local suppression is performed on the remaining unsafe combinations to remove them.

3.5.5 Cornell Anonymisation Toolkit (CAT)

In 2009 a group of students from Cornell University developed the Incognito Algorithm. Subsequently it was named the Cornell Anonymisation Toolkit (CAT) and made publicly available. Incognito implements a k-anonymity algorithm and takes into consideration all possible subsets of the indirect identifiers [71]. It was developed by Kristen LeFevre.

According to [69], given a data set that contains three quasi-identifiers or variables, for example age, date of birth, and gender, the Incognito algorithm would take into consideration the variables separately as well as a combination of the variables together. It then determines if any of these values identify unique or rare combinations

that could be explicitly identifying. Thereafter it uses optimisations to speed up its calculations. This allows for the practical application of larger data sets to be possible.

3.5.6 The University of Texas (Dallas) (UTD) Anonymisation Toolbox

The UTD Anonymisation Toolbox is an open source Java software tool that incorporates algorithms for k-anonymity and attribute disclosure control. In 2012, the most recent version was released [72] which incorporated a graphical user interface for researchers to easily arrange parameters of the available anonymisation algorithms. In addition, this version introduced an application programming interface (API) for developers. The new API enabled integration of anonymisation algorithms into various privacy-preserving data processing applications. As of the most recent release, the toolbox only supported unstructured text files (also called American Standard Code for Information Interchange [ASCII files]) which was an immediate disadvantage in comparison to the other tools available [73].

3.5.7 The sdcMicro package in R

The statistical disclosure control (sdcMicro) package provides some basic de-identification functions. sdcMicro was developed as a package for use within the R statistics software. The intention was for it to not run independently as a standalone application like other anonymisation tools [74]. Like other anonymisation tools, sdcMicro also contains a graphical user interface; however, functionality is limited. It only incorporates the k-anonymity and ℓ-diversity privacy models [5].

3.5.8 Anonymisation Tool Selection

Figure 6 below shows a comparison of the various tools mentioned in the previous section against the support, usability, and anonymity criteria [5]. Depending on the use case and users' requirements, the choice of a suitable tool differs.

		UTD-AT	CAT	sdcMicro	µ-Argus	ARX
Developer Support	Open source	Yes	Yes	Yes	No	Yes
	Active	No	No	Yes	No	Yes
	Public API	No	No	Yes	No	Yes
	Extensibility	Low	Low	Low	No	High
	Cross-platform	Yes	Yes	Yes	No	Yes
	Prog. Language	Java	C++	R	C++	Java
Usability	GUI coverage	None	Full	Partial	Full	Full
	Hierarchy creation	No	No	Yes	No	Yes
	Visualization	No	Data, Risks	Data, Risks	Risks	Data, solution space
	Data sources	CSV	Proprietary	CSV, Various	CSV, Various	CSV, Excel, DBMS
	Hierarchy format	Proprietary	Proprietary	Proprietary	Proprietary	CSV
	Standalone	No	Yes	No	Yes	Yes
Anonymity Methods	Automatic solution	Yes	Yes	Partial	No	Yes
	Privacy criteria	k, ℓ, t	ℓ, t	k, ℓ	None	k, ℓ, t, δ
	Generalization	Yes	Yes	Yes	Yes	Yes
	Tuple suppression	Partial	No	Yes	Yes	Yes
	Risk assessment	No	Limited	Yes	Yes	Limited

Figure 6: Comparison of anonymisation tools [5]

ARX was selected as the tool of choice for this study. A key reason for using ARX is that it is open source and freely available to all users. Another feature is that ARX can measure utility of a data set by comparing the results of different transformations, which was an important factor in considering the choice of tool. This functionality will allow for the measurement of information loss as a result of anonymisation. These two features are integral to evaluate the hypothesis proposed.

Furthermore, an outcome of the research was to measure the success of using a data anonymisation tool, specifically to ensure POPI-compliant data, as well as to determine whether the k-anonymity privacy model or l-diversity privacy model is the preferred model to use to achieve POPI compliance of a given customer data set. ARX supports both privacy models.

3.6 Data Privacy Measures

When trying to resolve data privacy issues, two parameters must be taken into consideration. Firstly, there is a need to measure data utility, and secondly, to estimate the privacy risk of the output data set after anonymisation. Attacker models can be used to estimate the privacy risks [75]. Both are discussed below.

3.6.1 Data Utility

When using data for data mining purposes, the process of privacy-preservation and transforming data can reduce data quality which in turn can lead to a loss of information [55]. Subsequently, this loss of information can also be attributed to a loss of utility when comparing the output data set to the original data set. However, a key trade-off when it comes to data utility is how to preserve as much utility as possible while still retaining privacy levels [1] .

The issue with utility-based privacy-preserving data mining was first studied formally in [40]. This study determined that the aim of privacy-preserving data publishing was to maximise “good utility” whilst at the same time reduce the ability for the identification of individuals in a data set. The study also proposed the separation of publishing marginal tables that have attributes which retain utility which are at the same time a problem for preserving and maintaining privacy [40].

Furthermore, in some studies, negative results related to dimensionality suggest that there needs to be a suppression of certain attributes to preserve privacy [7]. However, careful consideration must be taken to ensure that privacy preservation is done in a way to preserve utility. Therefore, it is important to understand what measures of utility are available so that the level of usefulness of the resultant output data set can be measured.

3.6.2 Utility Metrics

In this section, the available utility metrics that measure data utility are discussed. A brief description of each utility measure is also provided. In subsequent chapters, those utility measures that are in scope for this study are further highlighted.

The following important data utility measures are available:

- **Discernibility Metric (DM)** is based on the equivalence class size and involves introducing a penalty for suppressed records [76].

- **Average Equivalence Class Size (AECS)** measures the average size of groups of records that are unable to be identified and measures the equivalence class size. Average equivalence class size is also known as average class size [77][75].
- The **Precision** measure, introduced by Latanya Sweeney, summarises the level of generalisation applied to all attribute values [11].
- **Information Loss (IL)** was initially proposed by Iyengar et al. and measures the granularity of the data by determining the coverage of an attribute's domain that is contained in the transformed values [75] [78].
- **Non-Uniform Entropy** was first proposed by Gionis and Tassa [59]. It "computes a distance between the distribution of attribute values in an anonymised data set and the distribution of attribute values in the original data set" [79].
- **Classification Metric (CM)** applies when tuples are "assigned a categorical class label in an effort to produce anonymisations whose induced equivalence classes consist of tuples that are uniform with respect to the class label" [76]. Another definition of the Classification Metric is it is the sum of the individual penalties for each row in the table normalised by the total number of rows [78].

3.6.3 Attacker Models

When it comes to the measurement of privacy risks, attacker models are used. There are three different attacker models available [75].

The first attacker model is the prosecutor model. In this model the assumption is that the attacker is already aware that the respondent in part of the data set. The second attacker model is the journalist model where background information about the respondent is not known by the attacker. The third attacker model is the marketer model where the aim is to attack a larger number of individuals in a data set and not one individual only [41].

3.7 Summary

This chapter commenced with a discussion of the concepts of privacy models and anonymisation techniques and privacy threats. A brief overview of the various anonymisation tools was then outlined in order to provide context for the selection of a tool. Finally, data privacy regarding data utility and attacker models was discussed. In the next chapter, the design considerations for the study are provided.

4 DESIGN

Having shown the reader the numerous privacy models and techniques that can be used to perform data anonymisation in the previous chapter, a reference was made to the tools available which incorporate these privacy models as part of the software. The ARX anonymisation tool was selected in the last chapter. However, it must be mentioned that the approach is generic such that it can be used with any anonymisation tool of choice.

In this chapter, an overview of the research methodology and approach used in this study will be presented. This is crucial to validate the hypothesis outlined in the first chapter. Firstly, an examination of the characteristics, architecture, privacy models, and utility measures that are incorporated as part of the tool are presented. The data set and environment where the practical application of the anonymisation tool and the relevant setup of the environment is then explained to ensure a proper evaluation using a suitable approach to meet the aim of the study. Finally, a simple workflow to show the anonymisation process that will be followed is presented.

4.1 Research Methodology

In this section the research methodology carried out in this study is described. The methodology will show which method was used to test the hypothesis, the data collection and analysis thereof, as well as a justification on why the particular methodology was chosen.

4.1.1 Research Method

For this study, the technique chosen was to perform an experiment which is based within quantitative research design. Experiments are used to investigate causal relationships using tests controlled by the researcher [80]. In this study, an experiment was used to test the hypothesis where data anonymisation had to be carried out on a data set representing realistic customer data. As a result of the experiment, it was then possible to test the cause-and-effect relationship of the two privacy models implemented against each other with regards to privacy and data utility measures. To

do this, a data set was imported into a software tool, ARX, was configured for purposes of testing the hypothesis and thereafter evaluation of the results. The results of each privacy model were measured and reported on.

4.1.2 Data Collection and Analysis

Quantitative data was used in this study as hypothesis testing and understanding of anonymisation using privacy models were required [81]. The data set was obtained from the AdventureWorks 2016 Microsoft database that simulates a product sampling data warehouse supporting online transaction processing (OLTP) processing. Within the database resides information for a fictitious company called Adventure Works Cycles that manufactures bicycles [82].

In addition, data used in this study was specifically obtained from the customer (dim.customer) table within the AdventureWorks database. The reason this data was used as the sample was because it contains the structure, schema, and columns of a real-life customer database that a banking institution could have. Due to the nature of the testing performed, un-transformed and raw data was needed to accurately represent the sample population.

Below is an outline of the basic properties of the AdventureWorks data set used in this study:

- Data set Name: AdventureWorks 2016
- Data set Origination: Microsoft SQL sample data warehouse
- Data set Location: https://docs.microsoft.com/en-us/sql/samples/adventureworks-install-configure?view=sql-server-2017
- Data set Records: The complete data set contains 18,484 records.
- Data set File Size: The file sizes are 3095KB
- Data set File Format: Excel CSV file exported from a SQL database

In this study an important note to make is the distinction of the type of data that was used: the data was a static, structured, well-defined, textual, single-level data set. The

reason being is that for the POPI principle “Security Safeguards and Controls” there is a need for the data set that is being anonymised to be as close to real life data as possible, and which would typically reside in an organisations database[10]. Further definitions of these terms are described below to ensure clarity on the type of data that was tested [83]:

- **Static** refers to data that is not changing and is available completely when the anonymisation process is done. This is in contrast with streaming data where new data is constantly available. As a result, it is important to note that other anonymisation techniques may be needed for streaming data that is not described further in this study.

- **Structured** refers to data that is formatted properly and located in a known location.

- **Well-defined** relates to when the data set conformed to pre-defined rules as with the case of relational databases.

- **Textual** data refers to values that represent data that is alphanumeric in form, for example text, numbers, dates, etc. Anonymisation techniques for streaming data like audio, video, images, etc. create additional challenges and are not in the scope of this study.

- **Single-level** refers to data relating to various individuals with only one entry (and not multiple entries) per individual.

The AdventureWorks 2016 database backup file was downloaded from the link as mentioned above and installed in the SQL Server environment. The “dimcustomer” table was then exported to a CSV file. This resulted in 18,484 unique customer records.

The data set in total comprised 29 attributes. Table 3 refers to the entire data set. However, in this study the following 19 attributes from the dimcustomer data set were

used in the data anonymisation process. These are marked in the table below as “Yes”.

Table 3: Data set Attributes for dimcustomer Table

Number	Attribute	Included in Anonymisation
1	CustomerKey	Yes
2	GeographyKey	No
3	CustomerAlternateKey	No
4	Title	No
5	FirstName	Yes
6	MiddleName	Yes
7	LastName	Yes
8	NameStyle	No
9	BirthDate	Yes
10	MaritalStatus	Yes
11	Suffix	No
12	Gender	Yes
13	EmailAddress	Yes
14	YearlyIncome	Yes
15	TotalChildren	Yes
16	NumberChildrenAtHome	Yes
17	EnglishEducation	Yes
18	SpanishEducation	No
19	FrenchEducation	No
20	EnglishOccupation	Yes
21	SpanishOccupation	No
22	FrenchOccupation	No
23	HouseOwnerFlag	Yes
24	NumberCarsOwned	Yes
25	AddressLine1	Yes
26	AddressLine2	No

27	Phone	Yes
28	DateFirstPurchase	Yes
29	CommuteDistance	Yes

As shown above, the data set represented a realistic customer data set that could be used as the data required in this study. As the dataset contained information from the data warehouse sample scenario, the data was clear, concise and fit for purpose. There were 18,484 unique customer record which represented a sample size that was adequate for purposes of this study. Furthermore, the 19 attributes selected were based on characteristics that was needed to show the anonymisation of various types of identifiers e.g. string, integer, date/time as well as represent customer data that could require adherence to the POPI principle “Security Safeguards and Controls”.

4.1.3 Justification

An experiment was used primarily because the software tool selected could accurately compare the results of two anonymisation techniques within a closed environment using the same set of data and report the results empirically. An important consideration for using an experiment with a software tool built for the purpose of anonymising data was that the hypothesis could be tested in a practical way, which was important to the study, and therefore show meaningful results. The results could then also be represented in a tabular and graphic form, which would assist in showing the results of this study more clearly. Further details on the configuration of the software environment is shown later in this chapter.

4.2 ARX Characteristics

There are many commercial tools available that can be implemented which can transform a data set into an anonymised or de-identified data set [67]. Many of these tools are either open source or commercially available requiring a licence for use. A summary of these tools was provided in Chapter 3.

All experiments in this study were performed using the open source data anonymisation tool ARX, which was configured to use local generalisation. ARX allows the user to amend and change personally-identifying information (PII) data so that it can be shared. ARX is an anonymisation tool which has been developed specifically for the biomedical and health industry [42].

As the tool selected for this study plays a critical role in de-identifying a data set, the following advantages and disadvantages have been noted:

4.2.1 Advantages of ARX

- Contains a graphical user interface (GUI) that is easy to use with minimal training [66];
- Can visually guide the user step-by-step through the anonymisation process [66];
- Is highly configurable and scalable so that the implementation of anonymisation methods can be increased with a growth in future use cases as required [52];
- Independent API to be used separate to the GUI in a Java environment [5].

4.2.2 Disadvantages of ARX

- Knowledge of how to configure the various privacy models is required;
- Does not have functionality for data cleansing;
- Future work planned to incorporate differential privacy algorithms;
- Limited risk assessment models;
- The quasi-identifiers required in the k-anonymity and l-diversity privacy models are quite difficult to specify beforehand as it is not known what information is available to adversaries [39].

ARX has a comprehensive list of supported privacy models for anonymising structured data. Basic risk analysis and risk-based anonymisation can be performed within ARX. Syntactic privacy models are catered for including various methods that allow for the automatic and manual analysis of the utility of the data. A key goal of ARX is to obtain data sets that comply with syntactic privacy models whilst at the same time reduce attacks and prevent privacy breaches.

4.3 ARX Architecture

In this section a brief overview of the architecture developed within ARX is shown. As described in [5], the subsystems within the framework are tightly coupled to ensure extensibility. All the modules mentioned below can be used either with the API or the graphical user interface. Figure 7 below shows the full architecture layout of ARX, with a few key descriptions of the modules as well.

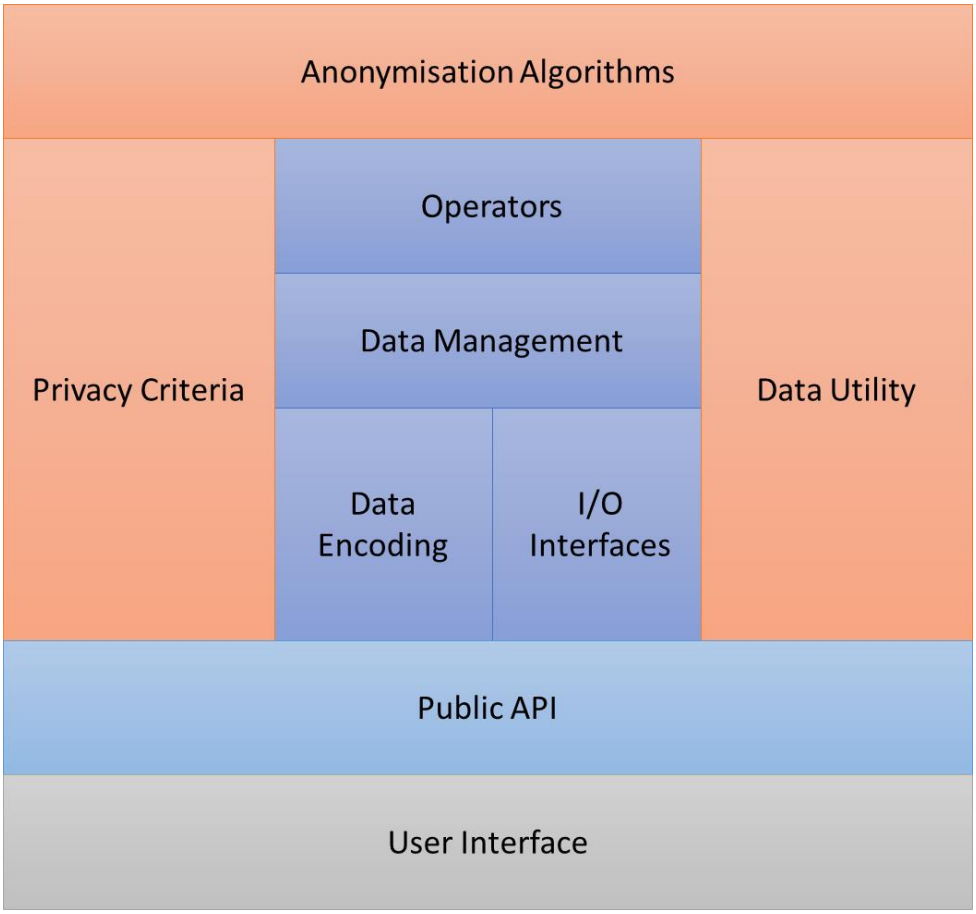


Figure 7: ARX High-Level Architecture Layout [5]

The following four core modules, which forms the basis of how ARX works, from [5] are:

1. The I/O modules, which enables the read/write of data to external drives;
2. The data encoding module, in which the data is transformed into the correct memory layout and format;

3. The data management module, that implements optimisation and internal representation of the data elements; and
4. The operators that deal with data record grouping and perform computations of frequency distribution over sensitive attributes.

The above are the core modules of ARX, however, the following extensible modules also form part of the tool:

1. Privacy criteria
2. Data utility
3. Anonymisation algorithms

The privacy criteria are implemented as well as the metrics for measuring data utility. Functionality for anonymisation algorithms to plug into the framework is also catered for.

4.4 ARX Privacy Models

In Section 3.3 Privacy Models, the list of the most common privacy models was discussed. In this section, a choice of privacy models that are available in ARX are selected for this study.

To review on previous work, ARX supports a variety of privacy models, namely syntactic privacy models, statistical privacy models as well as semantic privacy models [66]. Examples of specific models and types are shown in Table 4 below:

Table 4: Privacy Model Types [66]

Syntactic Privacy Models	**Statistical Privacy Models**	**Semantic Privacy Models**
k-anonymity	k-map	(ε, δ)-differential privacy
l-diversity	Thresholds on average risk	Game-theoretic de-identification approach

t-closeness	Methods based on super-population models	
δ-disclosure privacy		
β-likeness		
δ-presence		

In an effort to prevent data privacy breaches, syntactic privacy models are applied on a particular data set. In this study, the proposed hypothesis requires an evaluation of privacy models to ensure that the utility results and the privacy risks of the data set post-anonymisation can be measured.
To this end, in this study the following privacy models were selected:

- K-anonymity privacy model
- L-diversity privacy model

4.5 ARX Utility Measures

As mentioned in the previous chapter, in data anonymisation, utility measures are used to automatically compare data transformations to determine an optimal solution for the anonymisation whilst still maintaining utility.

Quality models are used to quantify the measurement for data quality [55]. In order to measure utility for this study, data quality models are implemented on the anonymised data. As a result, the data quality, and in turn data utility, is measured and reported.

The following attribute-level quality models are implemented within ARX in this study:

- Granularity
- Precision
- Squared Error
- Non-Uniform Entropy

The following data set-level quality models were applied in this study:

- Discernibility
- Ambiguity
- Average class size

- Record-level squared error

4.6 Experimental Environment

The experimental evaluation was setup on the following environment. The computing hardware was an Intel Core i5-5350U CPU @ 1.80GHz with 8GB of RAM installed.

The operating system installed was Windows 10 64-bit.

The anonymisation tool ARX Version 3.7.1 was installed. This version was released on 3 August 2018.

4.7 Data Identifiers

Within the ARX tool the various types of data identifiers are supported. These data identifiers are allocated to each attribute to allow for the appropriate anonymisation to be performed.

ARX supports the following attribute types with its relevant transformation method:

- Identifying attributes are the most obvious and will be taken out from the data set;
- Transformation will be applied to quasi-identifying attributes which together in some way are used to identify records;
- Sensitive attributes can be protected using privacy models, such as t-closeness or l-diversity or remain as-is and not transformed;
- Insensitive attributes are not changed in any way.

In this study, the following attribute types have been assigned to each attribute in the data set in Table 5 below:

Table 5: Attribute Types per Attribute

Number	Attribute	Attribute Type
1	CustomerKey	Identifying
2	FirstName	Identifying
3	MiddleName	Identifying
4	LastName	Identifying
5	BirthDate	Quasi-identifier
6	MaritalStatus	Quasi-identifier
7	Gender	Quasi-identifier
8	EmailAddress	Identifying
9	YearlyIncome	Quasi-identifier
10	TotalChildren	Quasi-identifier
11	NumberChildrenAtHome	Quasi-identifier
12	EnglishEducation	Quasi-identifier
13	EnglishOccupation	Sensitive
14	HouseOwnerFlag	Quasi-identifier
15	NumberCarsOwned	Quasi-identifier
16	AddressLine1	Identifying
17	Phone	Identifying
18	DateFirstPurchase	Quasi-identifier
19	CommuteDistance	Insensitive

4.8 Data Transformation Methods

4.8.1 Generalisation Hierarchies

Within ARX, generalisation hierarchies can be manually created or semi-automatically created. Common attribute types, such as numerical (discrete or continuous) and categorical variables can be created partially automatically. Values are grouped by a natural or user-defined method thereby creating hierarchies. These are mapped using user-defined intervals or created automatically through the wizard [5].

Generalisation hierarchies can be specified in two ways:

1. Using the built-in wizards
2. Importing and exporting hierarchy specifications

As the use of hierarchies is performed when using k-anonymity, further details on the various types of generalisation hierarchies are detailed below.

ARX supports four types of hierarchies when created automatically using the wizard:

- Masking-based hierarchies are used for various attribute types;
- Interval-based hierarchies are used for values with a ratio scale;
- Order-based hierarchies are used for values on an ordinal scale;
- Date-based hierarchies can be used for data ranges.

In this study, the following data transformation methods were applied to the attributes as shown in the Table 6 below. Where generalisation or suppression was used it was specifically mentioned below:

Table 6: Data Transformation Type per Attribute

Number	Attribute	Data Transformation
1	CustomerKey	Suppression
2	FirstName	Suppression
3	MiddleName	Suppression
4	LastName	Suppression
5	BirthDate	Date Based Generalisation Hierarchy
6	MaritalStatus	Generalisation
7	Gender	Generalisation
8	EmailAddress	Suppression
9	YearlyIncome	Generalisation
10	TotalChildren	Generalisation
11	NumberChildrenAtHome	Generalisation
12	EnglishEducation	Generalisation
13	EnglishOccupation	BLANK

14	HouseOwnerFlag	Generalisation
15	NumberCarsOwned	Generalisation
16	AddressLine1	Suppression
17	Phone	Suppression
18	DateFirstPurchase	Generalisation
19	CommuteDistance	None

4.9 Attribute Metadata

Attribute metadata relates to the types of data that a single attribute would represent. It is critical that the correct attribute data type is selected when using the built-in wizard for transformations. If not, inappropriate anonymisation could occur as a result.

ARX supports the following data types: string, integer, decimal, date/time, and ordinal.

In this study the following data types have been specified for the attributes in the data set shown in Table 7:

Table 7: Data Type per Attribute

Number	Attribute	Data Types
1	CustomerKey	Integer
2	FirstName	String
3	MiddleName	String
4	LastName	String
5	BirthDate	Date/Time
6	MaritalStatus	String
7	Gender	String
8	EmailAddress	String
9	YearlyIncome	Integer
10	TotalChildren	Integer
11	NumberChildrenAtHome	Integer
12	EnglishEducation	String

13	EnglishOccupation	String
14	HouseOwnerFlag	Integer
15	NumberCarsOwned	Integer
16	AddressLine1	String
17	Phone	String
18	DateFirstPurchase	Date/Time
19	CommuteDistance	String

4.9.1 Data Cleansing

Data imported into ARX cannot be changed; however, it does contain mechanisms to identify data quality issues by sorting the data, comparing, and analysing the data. A query can also be used to find records with data quality issues [5].

If there is a need for intense data clean up, this must be done outside of ARX and then imported for anonymisation to be performed.

In this study, the data set was not modified or cleansed to ensure that a realistic experiment sample was used.

4.10 Anonymisation Workflow

The steps to perform data anonymisation are available in ARX. These various steps are put together in a multi-step process which allows a user to adjust the parameters iteratively as required, until the output matches their need as shown in Figure 8 [5].

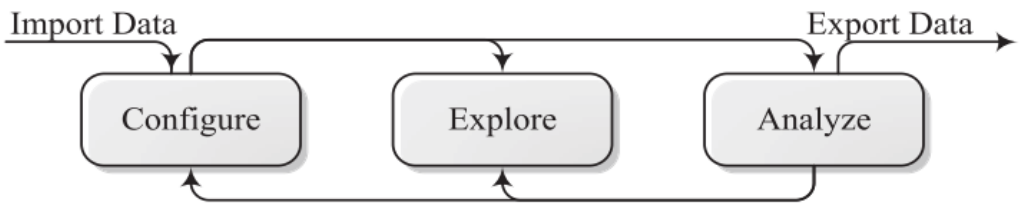


Figure 8: Anonymisation Workflow [5]

The three steps from the above anonymisation workflow are mapped to four perspectives in the ARX user interface [52] shown in Figure 9 below. This graphical user interface, using the perspectives as a guideline, is used to model different aspects of the anonymisation process end-to-end [66].

Figure 9: ARX Implementation Workflow Perspectives [66]

The workflow perspectives in Figure 9 above are used in this study to evaluate the hypothesis. In the following section a brief summary of each step, with its associated functionality, are described.

4.10.1 Configure

In the configuration phase, data is loaded, and the generalisation hierarchies are created or imported. Privacy models are selected and configured as well as utility measures specified. Lastly, transformation methods are configured [5]. After the configuration steps have been completed, the data can be anonymised.

4.10.2 Explore

In the exploration phase, the solution space (after the anonymisation has been completed) can be examined to look for data transformations that preserve privacy as well as meet the user needs [5]. An overview of possible solutions can be inspected in the exploration perspective. ARX also contains functionality to automatically propose a solution of choice.

4.10.3 Utility Analysis

In the analysis phase, a comparison of the input data as well as the transformed output data can be done. This is done so that the utility of the data can be assessed. Data utility is also analysed at this stage automatically using utility measures and metrics that are built into ARX [52]. Details of these utility metrics and utility measures were discussed in the previous chapter.

4.10.4 Risk Analysis

In the fourth perspective, different views of risk analysis on the output data can be seen. The risks associated with individual quasi-identifiers present in the data set as well as the distribution of class sizes are shown. Importantly, this view also displays details about estimated re-identification risks obtained from different models that are implemented when anonymising the data [52]. As a result, privacy risks can be analysed for an input data set as well as transformed output data.

In summary, based on the results of the various analyses that can be performed, the suitability of a solution candidate may either be confirmed, or the parameters of the anonymisation process can be modified to suit the user requirements, resulting in a semi-automated workflow.

4.11 Summary

As mentioned in this chapter, the approach and method for the testing of the hypothesis was presented. Details of the tool, ARX, was provided together with a view of the data set and its relevant attributes. The data transformation methods and workflows showed how the anonymisation process would be used to anonymise the data. In the next chapter the implementation of the design mentioned in this chapter is done.

5 IMPLEMENTATION

Having shown in the previous chapter the design of the anonymisation tool that was done, in this chapter these designs are implemented. Here the configuration of the anonymisation process steps are performed. Firstly, the setup of the environment is done to ensure an optimal testing environment.
Thereafter, seeing that ARX has a useful graphical user interface to guide the setup, the configuration of ARX is shown with screens of each step. Lastly the specific data anonymisation settings like privacy models, utility measures, and attribute properties are configured.

5.1 Environment Setup

5.1.1 AdventureWorks Data set

The AdventureWorks2016 data set was downloaded directly from the Microsoft site at the link below.

https://docs.microsoft.com/en-us/sql/samples/adventureworks-install-configure?view=sql-server-2017

The data warehouse file version AdventureWorksdW2016.bak was downloaded.

5.1.2 Database Server

Microsoft SQL Server 2016 was then installed in Evaluation mode. This was done so that the AdventureWorks backup database could be imported into the database server, and thereafter the customer information could be extracted.

The following SQL Server configuration was created in MS SQL Server 2016:

- Server Name: SAGREN
- Instance Name: MSC2019

5.1.3 Database Server Management Tool

SQL Server Management Studio (SSMS) version 2014 was installed as it is a tool to configure, monitor, and administer instances of SQL Server and databases. SSMS was used to restore the downloaded version of the AdventureWork2016 data warehouse. Default settings were selected upon installation to perform the restore.

Once the database was restored, all customer records from the dim.customer table within the AdventureWorks2016 database was extracted in an Excel file. Table 8 below shows the details of the dim.customer table.

Table 8: AdventureWorks Database Detail

AdventureWorks2016	
Database Name	AdventureWorks2016
Table	dim.customer
Rows	18,484
Columns	20

5.2 ARX Setup

ARX version 3.7.1 was downloaded from the product website https://arx.deidentifier.org/ and installed.
The default settings as proposed during the setup were chosen when installing ARX. Once installed, a new project was created and saved in preparation for the data set proposed in this study to be imported to.

5.3 Configuring ARX

In this section, the data set was loaded and the privacy models with the generalisation hierarchy were configured. The various transformation models are specified for the anonymisation process. Once the anonymisation process was completed the solution space can be viewed to organise transformations that have been automatically

proposed. Figure 10 below shows an overview of the steps in configuring the environment for the anonymisation to run. Each of these are described in the section below.

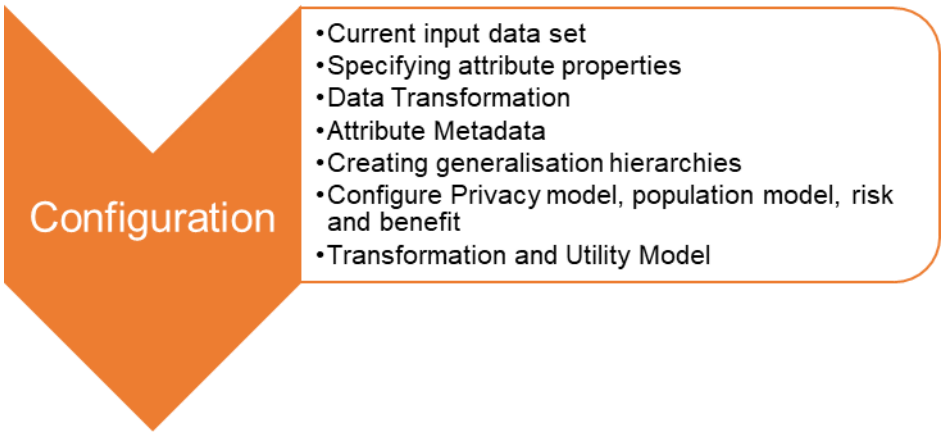
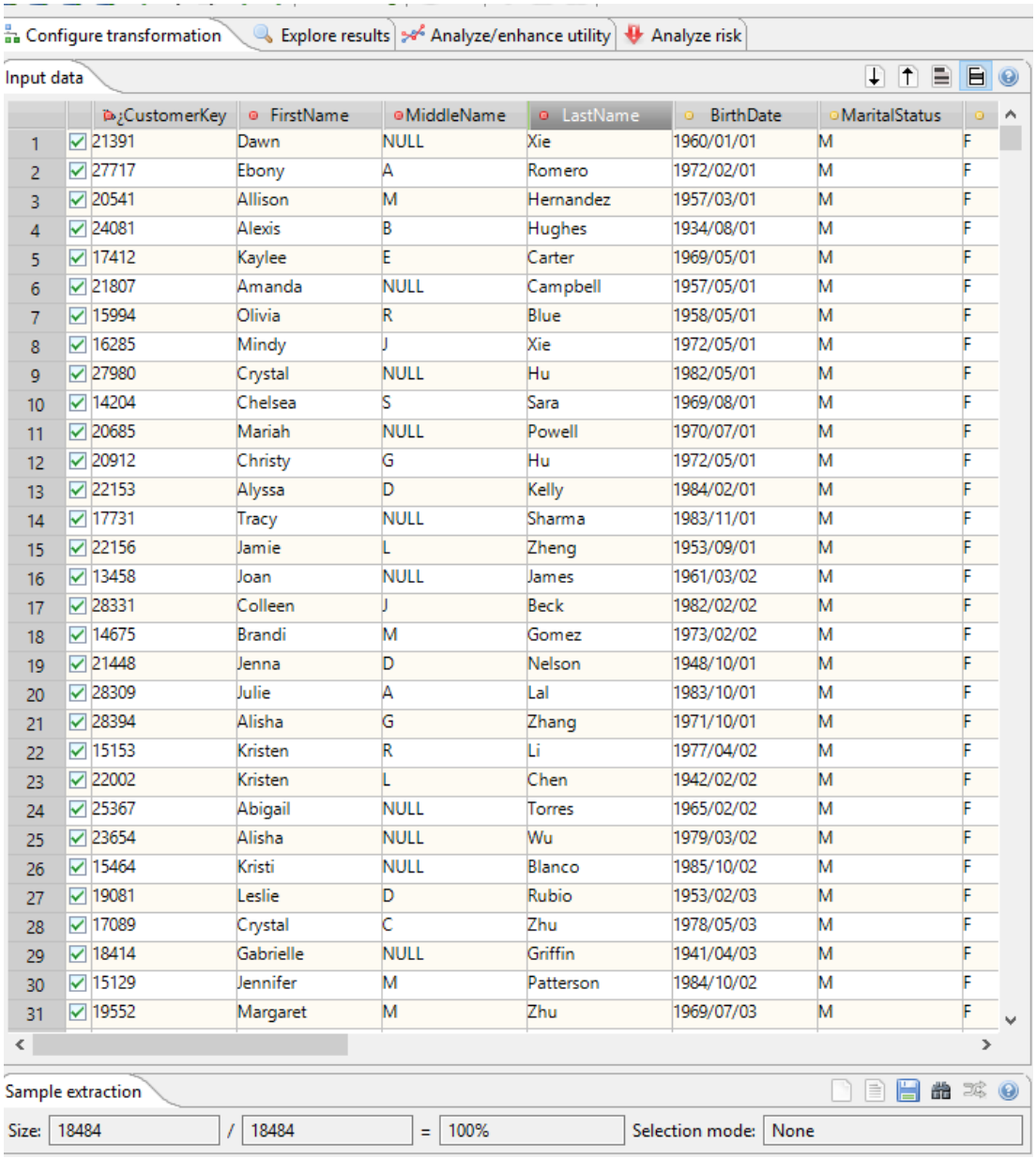


Figure 10: Configuration Perspective

5.3.1 Input Data Set

At this stage, the ARX tool was installed with a new project environment created. Furthermore, the data from the AdventureWorks2016 dim.customer table was imported using the built-in import wizard.

As shown in Figure 11 below, 18,484 records were imported.



	¿CustomerKey	FirstName	MiddleName	LastName	BirthDate	MaritalStatus	
1	21391	Dawn	NULL	Xie	1960/01/01	M	F
2	27717	Ebony	A	Romero	1972/02/01	M	F
3	20541	Allison	M	Hernandez	1957/03/01	M	F
4	24081	Alexis	B	Hughes	1934/08/01	M	F
5	17412	Kaylee	E	Carter	1969/05/01	M	F
6	21807	Amanda	NULL	Campbell	1957/05/01	M	F
7	15994	Olivia	R	Blue	1958/05/01	M	F
8	16285	Mindy	J	Xie	1972/05/01	M	F
9	27980	Crystal	NULL	Hu	1982/05/01	M	F
10	14204	Chelsea	S	Sara	1969/08/01	M	F
11	20685	Mariah	NULL	Powell	1970/07/01	M	F
12	20912	Christy	G	Hu	1972/05/01	M	F
13	22153	Alyssa	D	Kelly	1984/02/01	M	F
14	17731	Tracy	NULL	Sharma	1983/11/01	M	F
15	22156	Jamie	L	Zheng	1953/09/01	M	F
16	13458	Joan	NULL	James	1961/03/02	M	F
17	28331	Colleen	J	Beck	1982/02/02	M	F
18	14675	Brandi	M	Gomez	1973/02/02	M	F
19	21448	Jenna	D	Nelson	1948/10/01	M	F
20	28309	Julie	A	Lal	1983/10/01	M	F
21	28394	Alisha	G	Zhang	1971/10/01	M	F
22	15153	Kristen	R	Li	1977/04/02	M	F
23	22002	Kristen	L	Chen	1942/02/02	M	F
24	25367	Abigail	NULL	Torres	1965/02/02	M	F
25	23654	Alisha	NULL	Wu	1979/03/02	M	F
26	15464	Kristi	NULL	Blanco	1985/10/02	M	F
27	19081	Leslie	D	Rubio	1953/02/03	M	F
28	17089	Crystal	C	Zhu	1978/05/03	M	F
29	18414	Gabrielle	NULL	Griffin	1941/04/03	M	F
30	15129	Jennifer	M	Patterson	1984/10/02	M	F
31	19552	Margaret	M	Zhu	1969/07/03	M	F

Figure 11: View of imported records in the data set

The following columns (attributes) from the dim.customer table were imported and are shown in Table 9:

Table 9: ARX Input Data Attributes

Number	Attribute
1	CustomerKey
2	FirstName
3	MiddleName
4	LastName
5	BirthDate
6	MaritalStatus
7	Gender
8	EmailAddress
9	YearlyIncome
10	TotalChildren
11	NumberChildrenAtHome
12	EnglishEducation
13	EnglishOccupation
14	HouseOwnerFlag
15	NumberCarsOwned
16	AddressLine1
17	Phone
18	DateFirstPurchase
19	CommuteDistance

The following ten attributes were not imported. This is because these columns either contain null values or were not appropriate for consideration when sampling realistic customer data.

1. GeographyKey
2. CustomerAlternateKey
3. Title
4. NameStyle
5. Suffix
6. SpanishEducation

7. FrenchEducation
8. SpanishOccupation
9. FrenchOccupation
10. AddressLine2

Specific records within a data set can also be selected for anonymisation, therefore, it specifies which records are contained in the project sample. However, for this study, and as shown in Table 9 above, all 18,484 records were imported.

5.3.2 Specifying Attribute Properties: Attribute Metadata

After the data set has been imported, the attribute metadata was configured. This is where the datatype for each attribute is set together with the format. Figure 12 below shows the configuration of each attribute datatype in the data set.

Data transformation | Attribute metadata

Attribute	Data type	Format
ï»¿CustomerKey	Integer	Default
FirstName	String	Default
MiddleName	String	Default
LastName	String	Default
BirthDate	Date/Time	dd/MM/yyyy (EN)
MaritalStatus	String	Default
Gender	String	Default
EmailAddress	String	Default
YearlyIncome	Integer	Default
TotalChildren	Integer	Default
NumberChildrenAtHome	Integer	Default
EnglishEducation	String	Default
EnglishOccupation	String	Default
HouseOwnerFlag	Integer	Default
NumberCarsOwned	Integer	Default
AddressLine1	String	Default
Phone	String	Default
DateFirstPurchase	Date/Time	dd/MM/yyyy (EN)
CommuteDistance	String	Default

Figure 12: Configuration of Attribute Metadata

5.3.3 Specifying Attribute Properties: Data Transformation

The type of an attribute can be set within the “Data Transformation” tab. The type of an attribute can be specified together with the transformation method to be applied to the data.

5.3.3.1 Attribute Type

To recap from the previous chapter, the data will be transformed according to the rules for the attribute types as below:

- Identifying attributes are the most obvious and will be taken out from the data set;
- Transformation will be applied to quasi-identifying attributes which together in some way is used to identify records;
- Sensitive attributes can be protected using privacy models, such as t-closeness or l-diversity or remain as-is and not transformed;
- Insensitive attributes are not changed in any way.

As shown below in Table 10, the following attribute types were set for each field:

Table 10: Configuration of Attribute Types

Number	Attribute	Attribute Type
1	CustomerKey	Identifying
2	FirstName	Identifying
3	MiddleName	Identifying
4	LastName	Identifying
5	BirthDate	Quasi-identifier
6	MaritalStatus	Quasi-identifier
7	Gender	Quasi-identifier
8	EmailAddress	Identifying
9	YearlyIncome	Quasi-identifier
10	TotalChildren	Quasi-identifier
11	NumberChildrenAtHome	Quasi-identifier
12	EnglishEducation	Quasi-identifier

13	EnglishOccupation	Sensitive
14	HouseOwnerFlag	Quasi-identifier
15	NumberCarsOwned	Quasi-identifier
16	AddressLine1	Identifying
17	Phone	Identifying
18	DateFirstPurchase	Quasi-identifier
19	CommuteDistance	InSensitive

5.3.3.2 Attribute Transformation

ARX includes the following transformation methods: generalisation, micro aggregation, and suppression. Attribute types that are identifying are automatically supressed. Attribute types that are quasi-identifiers have been generalised using generalisation hierarchies.

The full details of the data transformation type per attribute are shown in Table 11.

Table 11: Configuration of Attribute Transformation

Number	Attribute	Data Transformation
1	CustomerKey	Suppression
2	FirstName	Suppression
3	MiddleName	Suppression
4	LastName	Suppression
5	BirthDate	Date Based Generalisation Hierarchy
6	MaritalStatus	Generalisation
7	Gender	Generalisation
8	EmailAddress	Suppression
9	YearlyIncome	Generalisation
10	TotalChildren	Generalisation
11	NumberChildrenAtHome	Generalisation
12	EnglishEducation	Generalisation

13	EnglishOccupation	Blank
14	HouseOwnerFlag	Generalisation
15	NumberCarsOwned	Generalisation
16	AddressLine1	Suppression
17	Phone	Suppression
18	DateFirstPurchase	Generalisation
19	CommuteDistance	None

5.3.4 Configure Privacy Models

In this section, the configuration of the two privacy models, namely k-anonymity and l-diversity, are shown.

5.3.4.1 K-anonymity Configuration

For the project, the first privacy model k-anonymity was selected from the list of available privacy models. Within the configuration of k-anonymity, the value for k must be specified. This study was configured for 2 anonymity where k = 2. Refer to Figure 13 for the screen detail.

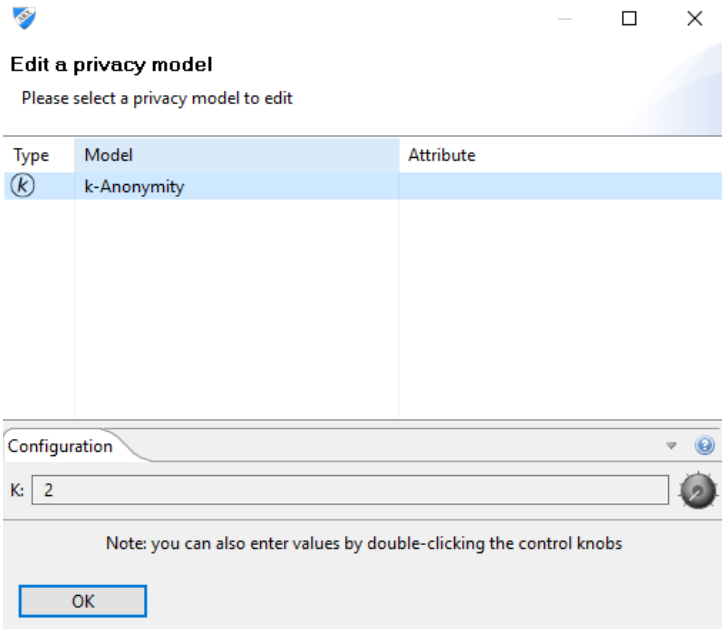


Figure 13: K-anonymity Privacy Model Configuration

5.3.4.2 L-diversity Configuration

The l-diversity privacy model was thereafter configured as shown in Figure 14. The attribute “EnglishOccupation” was defined to be a sensitive attribute when the transformation methods were selected. In this screen the attribute “EnglishOccupation” was selected with the property as “sensitive attribute”.

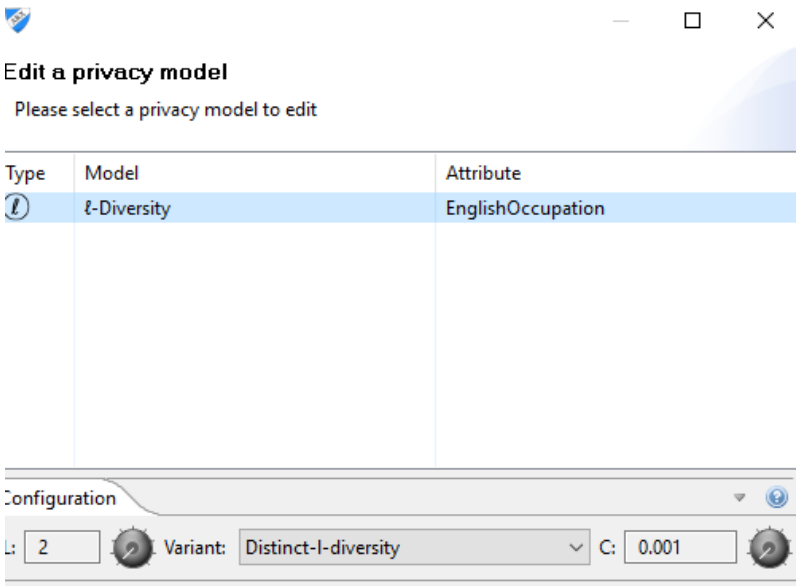


Figure 14 : L-diversity Privacy Model Configuration

5.3.5 Configure General Settings

Within the general setting tab, various general configurations can be done. Refer to Figure 15 below.

Firstly, the maximum number of outliers that may automatically be removed from the data set are specified. This is done by defining the suppression limit. The suppression limit is the maximum number of records that are removed. If the user selects 100% there will be no outliers in the data set. For this study, 100% was selected as the setting.

The recommended setting for the option "Approximate" is "off".

Precomputation reduces the execution times for certain utility measures. In this study the recommended setting is "off".

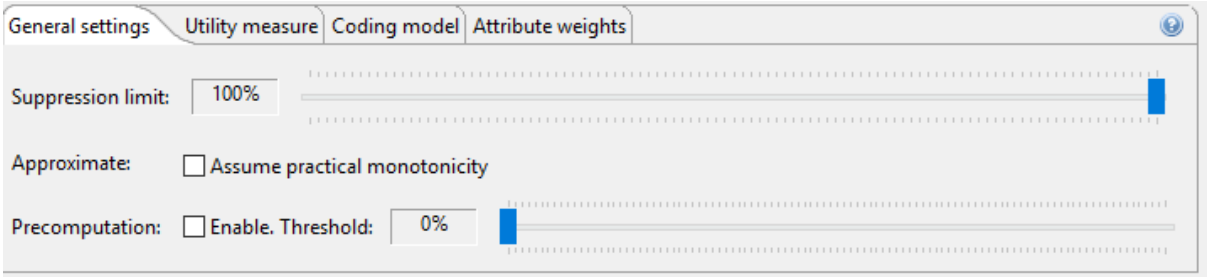


Figure 15: General Setting Configuration

5.3.6 Specify Utility Measures

In this tab, the utility measures that will be used in the study are configured. Refer to Figure 16 below.

Firstly, the measure for utility is selected. This is done by specifying the data quality model. This can be used as an optimisation function during the anonymisation process. The measure selected was “Loss”.

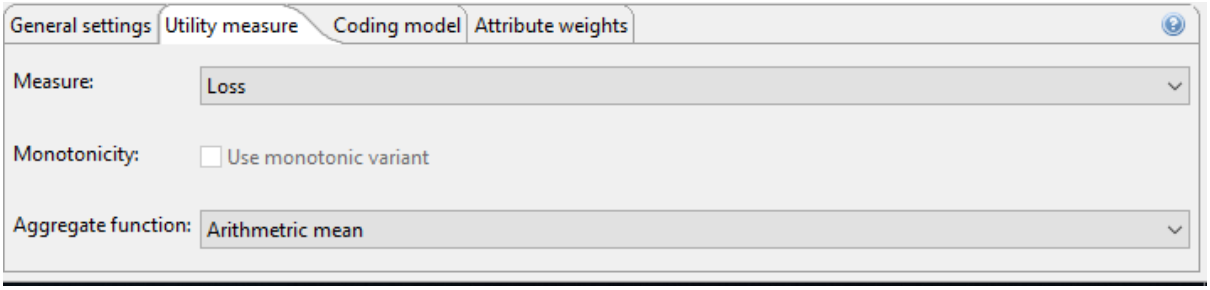


Figure 16: Utility Measure Configuration

Secondly, monotonicity can be used to make the anonymisation process more efficient. This, however, can lead to significant reductions in output data quality. Thus, in this study the recommended setting is "off".

Finally, user-defined aggregate functions are configured. These aggregate functions are used to estimate the individual attributes within a data set into a global value. For this study, the recommended setting is "Arithmetic Mean".

5.3.7 Specify Coding Model

There are quality models which inform whether generalisation or suppression should take preference when transforming data. With this option, more generalisation or more suppression can be selected. The default option selected for both generalisation and suppression was equally suited at 50%. Figure 17 below refers.

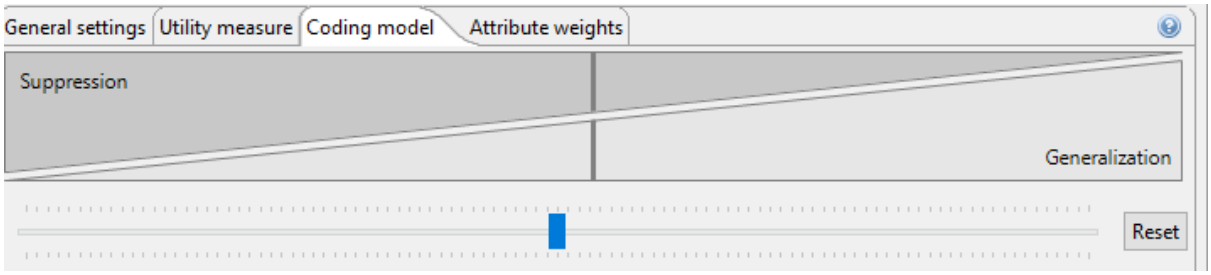


Figure 17: Coding Model Configuration

5.3.8 Specify Attribute Weights

Most privacy models support weighting an attribute to specify their level of importance in relation to other attributes. Therefore, when anonymising a data set, this functionality can be used to lessen the loss of information by assigning higher weights to certain attributes.

Weights are then assigned to various attributes to influence the level of information loss i.e. Gender can have more influence than Marital Status within a data set when taking into consideration the full set of attributes. However, in this study the default option of 0.5 was selected as a weighting for all attributes. Refer to Figure 18 below.

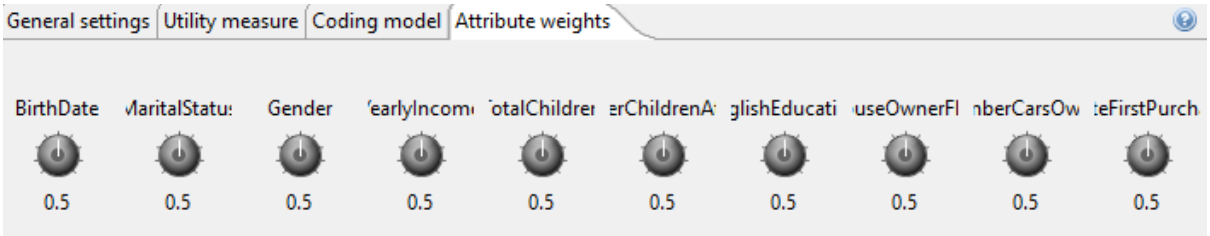


Figure 18: Attribute Weight Configuration

5.4 Summary

In this chapter the series of steps to configure the environment in a suitable way to ensure the testing of the hypothesis was done. The setup of the privacy models, attributes, and data transformations were done, with general settings also configured. The graphical user interface was useful in doing the above. In the next chapter the results of the implementation are provided.

6 TESTING/RESULTS

In the previous chapter, the anonymisation process was setup and the testing performed. The tests were run in a configured environment for the data anonymisation using the two privacy models selected. In this chapter, the testing phase of the implementation is examined. The results of the data anonymisation process using the two chosen privacy models is reported. The visualisation of the solution space is presented in a useful way showing the transformations results. The outcomes for the data utility, the data quality, and privacy risk measures are documented.

6.1 Exploring the Solution Space

Once the anonymisation process has been completed the solution space can be viewed to organise transformations that have been automatically proposed. Figure 19 shows a summary of the steps that can be performed within the software.

Figure 19: Exploration Considerations

6.1.1 Solution Space Testing

The first step after the anonymisation is applied is to view the outcome of the anonymisation. Visually the results of the transformations that have been applied to the data set can be seen in multiple views. This view is called the solution space. Therefore, the solution space is a graphical representation of the transformations that have occurred as a result of anonymisation. The solution space will show

transformations that fulfil the defined privacy criteria according to the model selected. Transformations are automatically selected based on improved data utility and transformations that support optimal data utility.

6.1.2 K-Anonymity: Solution Space Results

In this section the solution space results for the k-anonymity privacy model are shown. A Hasse diagram, by definition, is a “mathematical diagram that represents a finite partially ordered set, in the form of a drawing of its transitive reduction.” Figure 20 below shows a Hasse diagram of the underlying generalisation lattice. In this diagram, every node represents a single transformation. These are identified by the generalisation levels that are specified for the quasi-identifiers that occur in the data set.

Transformations are reflected using three background colours:

- Green: a transformation which results in a privacy-preserving data set.
- Red: a transformation which does not result in a privacy-preserving data set.
- Orange: the optimal transformation regarding a specified utility measure.

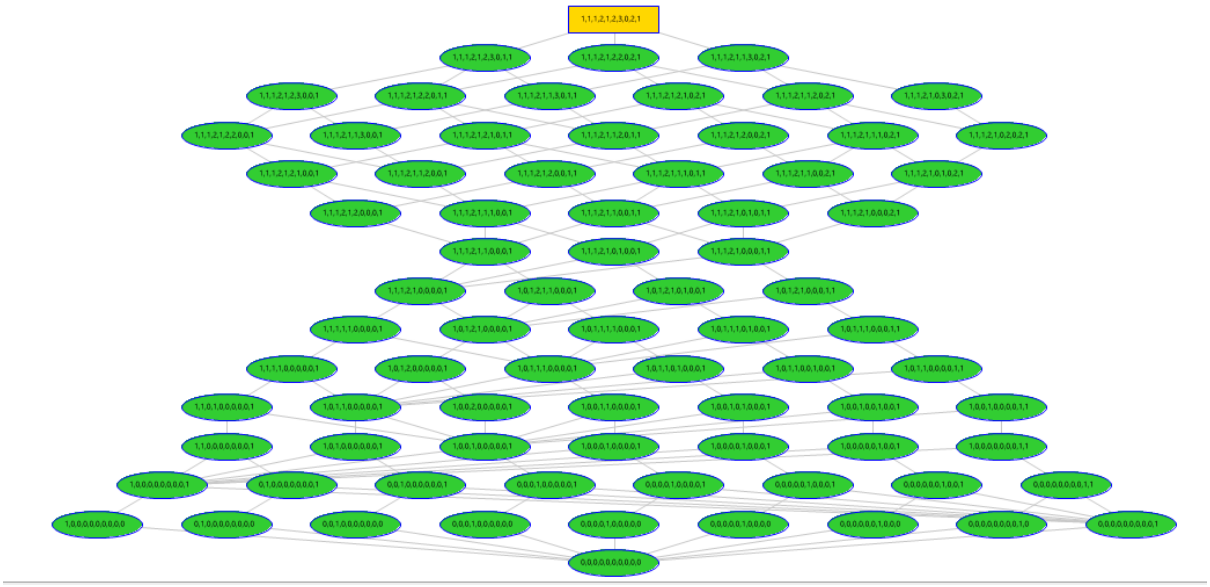
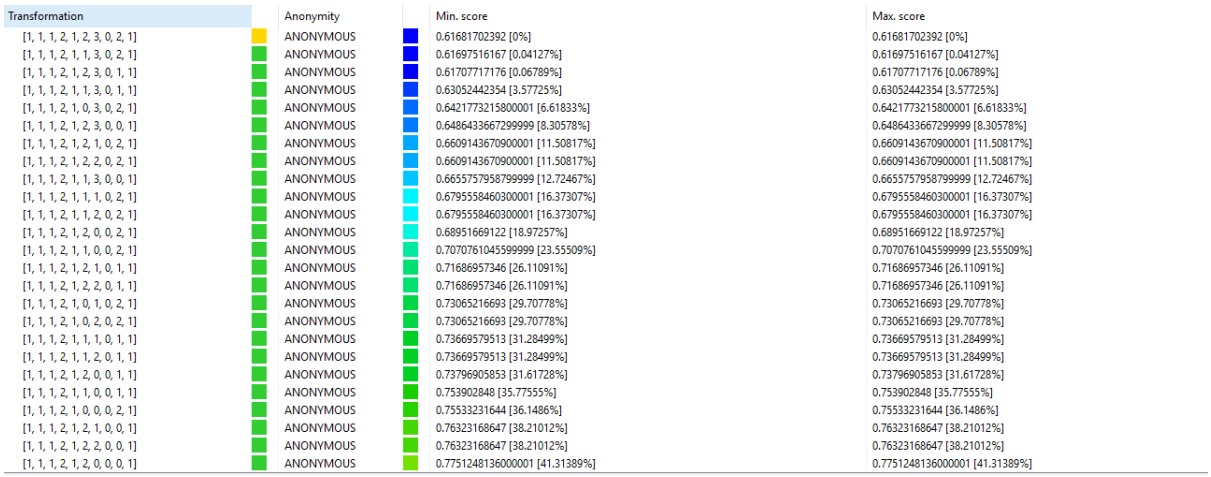


Figure 20: K-anonymity Solution Space Hasse Diagram View

Furthermore, the very same transformations as shown in a Hasse diagram in Figure 20 can be shown as a list view in Figure 21 or as a tile view in Figure 22 below.



Transformation	Anonymity	Min. score	Max. score
[1, 1, 1, 2, 1, 2, 3, 0, 2, 1]	ANONYMOUS	0.61681702392 [0%]	0.61681702392 [0%]
[1, 1, 1, 2, 1, 1, 3, 0, 2, 1]	ANONYMOUS	0.61697516167 [0.04127%]	0.61697516167 [0.04127%]
[1, 1, 1, 2, 1, 2, 3, 0, 1, 1]	ANONYMOUS	0.61707717176 [0.06789%]	0.61707717176 [0.06789%]
[1, 1, 1, 2, 1, 1, 3, 0, 1, 1]	ANONYMOUS	0.63052442354 [3.57725%]	0.63052442354 [3.57725%]
[1, 1, 1, 2, 1, 0, 3, 0, 2, 1]	ANONYMOUS	0.6421773215800001 [6.61833%]	0.6421773215800001 [6.61833%]
[1, 1, 1, 2, 1, 2, 3, 0, 0, 1]	ANONYMOUS	0.6486433667299999 [8.30578%]	0.6486433667299999 [8.30578%]
[1, 1, 1, 2, 1, 2, 1, 0, 2, 1]	ANONYMOUS	0.6609143670900001 [11.50817%]	0.6609143670900001 [11.50817%]
[1, 1, 1, 2, 1, 2, 2, 0, 2, 1]	ANONYMOUS	0.6609143670900001 [11.50817%]	0.6609143670900001 [11.50817%]
[1, 1, 1, 2, 1, 1, 3, 0, 0, 1]	ANONYMOUS	0.6655757958799999 [12.72467%]	0.6655757958799999 [12.72467%]
[1, 1, 1, 2, 1, 1, 1, 0, 2, 1]	ANONYMOUS	0.6795558460300001 [16.37307%]	0.6795558460300001 [16.37307%]
[1, 1, 1, 2, 1, 1, 2, 0, 2, 1]	ANONYMOUS	0.6795558460300001 [16.37307%]	0.6795558460300001 [16.37307%]
[1, 1, 1, 2, 1, 2, 0, 0, 2, 1]	ANONYMOUS	0.68951669122 [18.97257%]	0.68951669122 [18.97257%]
[1, 1, 1, 2, 1, 1, 0, 0, 2, 1]	ANONYMOUS	0.7070761045599999 [23.55509%]	0.7070761045599999 [23.55509%]
[1, 1, 1, 2, 1, 2, 1, 0, 1, 1]	ANONYMOUS	0.71686957346 [26.11091%]	0.71686957346 [26.11091%]
[1, 1, 1, 2, 1, 2, 2, 0, 1, 1]	ANONYMOUS	0.71686957346 [26.11091%]	0.71686957346 [26.11091%]
[1, 1, 1, 2, 1, 0, 1, 0, 2, 1]	ANONYMOUS	0.73065216693 [29.70778%]	0.73065216693 [29.70778%]
[1, 1, 1, 2, 1, 0, 2, 0, 2, 1]	ANONYMOUS	0.73065216693 [29.70778%]	0.73065216693 [29.70778%]
[1, 1, 1, 2, 1, 1, 1, 0, 1, 1]	ANONYMOUS	0.73669579513 [31.28499%]	0.73669579513 [31.28499%]
[1, 1, 1, 2, 1, 1, 2, 0, 1, 1]	ANONYMOUS	0.73669579513 [31.28499%]	0.73669579513 [31.28499%]
[1, 1, 1, 2, 1, 2, 0, 0, 1, 1]	ANONYMOUS	0.73796905853 [31.61728%]	0.73796905853 [31.61728%]
[1, 1, 1, 2, 1, 1, 0, 0, 1, 1]	ANONYMOUS	0.753902848 [35.77555%]	0.753902848 [35.77555%]
[1, 1, 1, 2, 1, 0, 0, 0, 2, 1]	ANONYMOUS	0.75533231644 [36.1486%]	0.75533231644 [36.1486%]
[1, 1, 1, 2, 1, 2, 1, 0, 0, 1]	ANONYMOUS	0.76323168647 [38.21012%]	0.76323168647 [38.21012%]
[1, 1, 1, 2, 1, 2, 2, 0, 0, 1]	ANONYMOUS	0.76323168647 [38.21012%]	0.76323168647 [38.21012%]
[1, 1, 1, 2, 1, 2, 0, 0, 0, 1]	ANONYMOUS	0.7751248136000001 [41.31389%]	0.7751248136000001 [41.31389%]

Figure 21: K-anonymity Solution Space List View

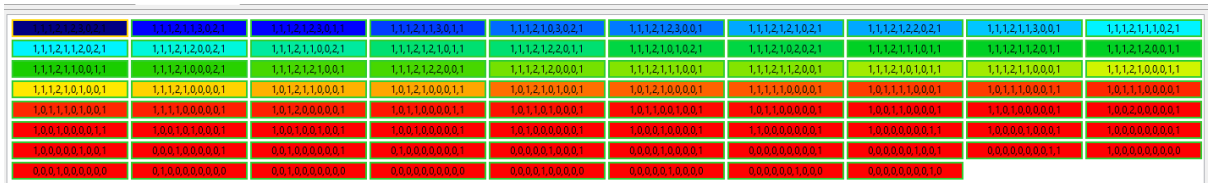


Figure 22: K-anonymity Solution Space Tile View

As each node, row, or set of tiles in the two figures above represents one transformation that is applied to the input data set, it can be clearly seen that 36,864 transformations have been applied to the data set using the k-anonymity privacy model.

6.1.3 L-Diversity: Solution Space Results

The solution space results for the l-diversity privacy model are shown in this section. Similar to the Hasse diagram view of the previous k-anonymity privacy model, Figure 23, Figure 24, and Figure 25 below shows the solution space results for the l-diversity privacy model in the respective views.

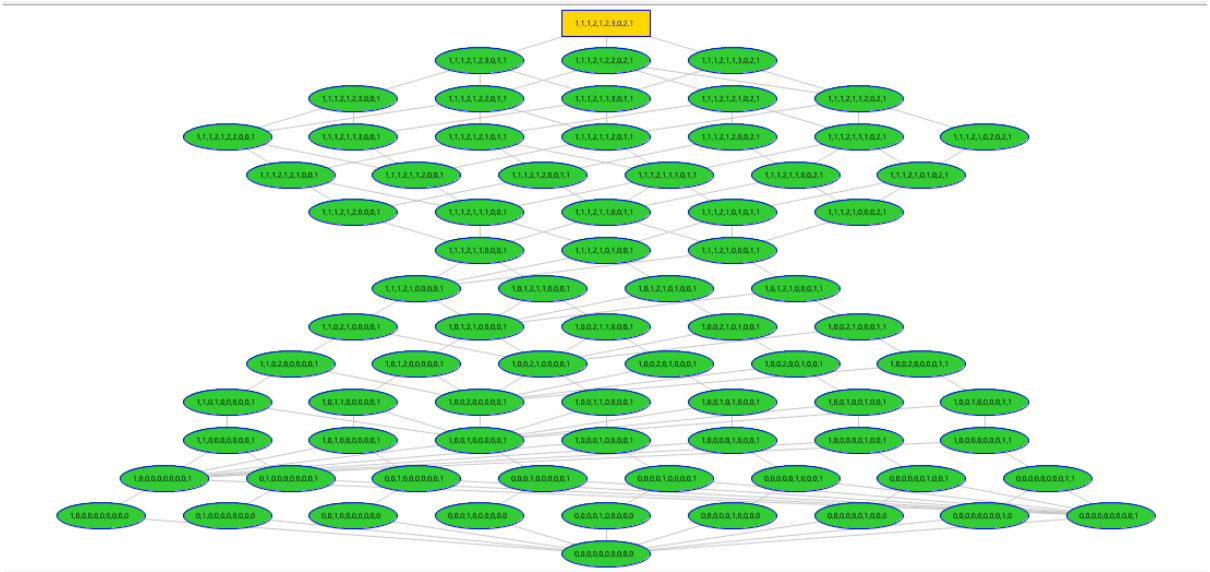


Figure 23: L-diversity Solution Space Lattice View

Transformation	Anonymity	Min. score	Max. score
[1, 1, 2, 1, 2, 3, 0, 2, 1]	ANONYMOUS	0.6368075119 [0%]	0.6368075119 [0%]
[1, 1, 1, 2, 1, 1, 3, 0, 2, 1]	ANONYMOUS	0.64522792141 [2.31844%]	0.64522792141 [2.31844%]
[1, 1, 1, 2, 1, 2, 3, 0, 1, 1]	ANONYMOUS	0.6536335821800001 [4.63282%]	0.6536335821800001 [4.63282%]
[1, 1, 1, 2, 1, 1, 3, 0, 1, 1]	ANONYMOUS	0.67358506971 [10.12619%]	0.67358506971 [10.12619%]
[1, 1, 1, 2, 1, 2, 3, 0, 0, 1]	ANONYMOUS	0.6920111395800002 [15.19955%]	0.6920111395800002 [15.19955%]
[1, 1, 1, 2, 1, 1, 3, 0, 0, 1]	ANONYMOUS	0.7148267281099999 [21.48151%]	0.7148267281099999 [21.48151%]
[1, 1, 1, 2, 1, 2, 1, 0, 2, 1]	ANONYMOUS	0.7197759710600001 [22.84421%]	0.7197759710600001 [22.84421%]
[1, 1, 1, 2, 1, 2, 2, 0, 2, 1]	ANONYMOUS	0.7197759710600001 [22.84421%]	0.7197759710600001 [22.84421%]
[1, 1, 1, 2, 1, 1, 1, 0, 2, 1]	ANONYMOUS	0.74511938702 [29.82217%]	0.74511938702 [29.82217%]
[1, 1, 1, 2, 1, 1, 2, 0, 2, 1]	ANONYMOUS	0.74511938702 [29.82217%]	0.74511938702 [29.82217%]
[1, 1, 1, 2, 1, 2, 0, 0, 2, 1]	ANONYMOUS	0.7526406637299998 [31.89305%]	0.7526406637299998 [31.89305%]
[1, 1, 1, 2, 1, 1, 0, 0, 2, 1]	ANONYMOUS	0.77589580864 [38.29603%]	0.77589580864 [38.29603%]
[1, 1, 1, 2, 1, 2, 1, 0, 1, 1]	ANONYMOUS	0.7804645472399998 [39.55397%]	0.7804645472399998 [39.55397%]
[1, 1, 1, 2, 1, 2, 2, 0, 1, 1]	ANONYMOUS	0.7804645472399998 [39.55397%]	0.7804645472399998 [39.55397%]
[1, 1, 1, 2, 1, 0, 1, 0, 2, 1]	ANONYMOUS	0.80299855762 [45.75839%]	0.80299855762 [45.75839%]
[1, 1, 1, 2, 1, 0, 2, 0, 2, 1]	ANONYMOUS	0.80299855762 [45.75839%]	0.80299855762 [45.75839%]
[1, 1, 1, 2, 1, 1, 1, 0, 1, 1]	ANONYMOUS	0.8051788335000001 [46.3587%]	0.8051788335000001 [46.3587%]
[1, 1, 1, 2, 1, 1, 2, 0, 1, 1]	ANONYMOUS	0.8051788335000001 [46.3587%]	0.8051788335000001 [46.3587%]
[1, 1, 1, 2, 1, 2, 0, 0, 1, 1]	ANONYMOUS	0.80659393161 [46.74833%]	0.80659393161 [46.74833%]
[1, 1, 1, 2, 1, 1, 0, 0, 1, 1]	ANONYMOUS	0.8253448965100001 [51.91115%]	0.8253448965100001 [51.91115%]
[1, 1, 1, 2, 1, 2, 1, 0, 0, 1]	ANONYMOUS	0.82748096257 [52.49928%]	0.82748096257 [52.49928%]
[1, 1, 1, 2, 1, 2, 2, 0, 0, 1]	ANONYMOUS	0.82748096257 [52.49928%]	0.82748096257 [52.49928%]
[1, 1, 1, 2, 1, 0, 0, 0, 2, 1]	ANONYMOUS	0.83007378643 [53.21318%]	0.83007378643 [53.21318%]
[1, 1, 1, 2, 1, 2, 0, 0, 0, 1]	ANONYMOUS	0.8429170888500002 [56.74941%]	0.8429170888500002 [56.74941%]
[1, 1, 1, 2, 1, 1, 1, 0, 0, 1]	ANONYMOUS	0.8466110043099999 [57.76647%]	0.8466110043099999 [57.76647%]
[1, 1, 1, 2, 1, 1, 2, 0, 0, 1]	ANONYMOUS	0.8466110043099999 [57.76647%]	0.8466110043099999 [57.76647%]

Figure 24: L-diversity Solution Space List View

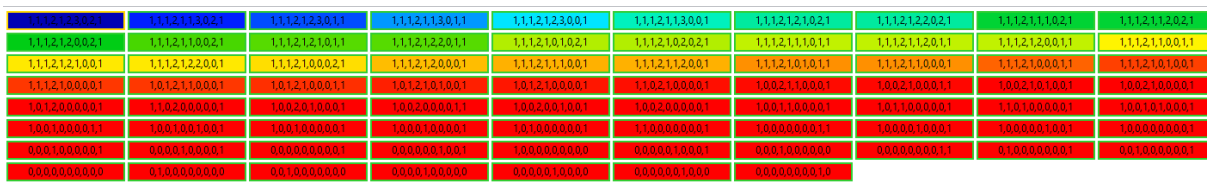


Figure 25: L-diversity Solution Space Tile View

It can be noted that the total number of transformations applied to the data set is 36,864 when using the l-diversity privacy model for anonymisation.

A further discussion on the analysis of the solution space results is done in the next chapter.

6.2 Analysing Data Utility

The quality and the utility of the output data after the anonymisation has been completed is reported in this section. Utility, in terms of individual attributes as well as the entire data set, is analysed and shown graphically next to each other to allow for easy comparison. Refer to Figure 26 below.

6.2.1 Data Utility Testing

Figure 26: Utility Analysis Considerations

When testing the utility of the resultant data set, the following features are analysed:

1. Data utility is analysed;
2. Input and output data are compared, i.e. the transformed data set is compared to the input data set;
3. Summary statistics of any selected attribute are shown;
4. Frequency distribution for values of individual attributes are shown;
5. A contingency heat map showing the contingency between two selected attributes;
6. Summary information for the equivalence classes and records in the data set are shown;

7. Properties of input and output data provides a basic display of the properties relating to the input data set;
8. Classification performance;
9. Data quality models are analysed.

For the context of this study, a key feature from the list above that will be discussed in more detail below is an analysis of the data quality models. The data quality models are implemented to measure the utility of the output data set.

6.2.1.1 Data Quality Models

Data quality models are used to display the measurements of the data quality outputs when using various general-purpose models. The results within the software show two different types of data quality, namely:

- Attribute-level quality shows measures relating to each quasi-identifier
- Data-level quality shows quality measures for the entire set of quasi-identifiers,

The following attribute-level quality models are implemented within ARX in this study [59]:

- Granularity
- Precision
- Squared Error
- Non-Uniform Entropy

Additionally, the following data set level quality models were applied in this study:

- Discernibility
- Ambiguity
- Average class size
- Record-level squared error

The data quality model results for each privacy model implemented are shown in detail in the next section.

6.2.2 K-anonymity Data Utility Results

In this section, the data quality obtained for output data for the k-anonymity privacy model is shown. Attribute-level as well as data set level quality results are displayed.

6.2.2.1 Data Quality Models

In Figure 27 below, the results of the testing for data quality when using the k-anonymity privacy model are shown.

Output data | Classification performance | Quality models

Attribute-level quality

Attribute	Data type	Missings	Gen. intensity	Granularity	N.-U. entropy	Squared error
BirthDate	String	17.70721%	41.1464%	79.45998%	32.74223%	81.65856%
MaritalStatus	String	100%	0%	0%	0%	0%
Gender	String	100%	0%	0%	0%	0%
YearlyIncome	String	17.70721%	49.72409%	38.98687%	10.25579%	64.82261%
TotalChildren	String	17.70721%	41.1464%	49.37568%	29.61031%	67.26952%
NumberChildr...	String	17.70721%	28.44803%	34.13763%	17.61558%	54.2183%
EnglishEducati...	String	100%	0%	0%	0%	0%
HouseOwnerFl...	Integer	17.70721%	82.29279%	82.29279%	81.08588%	82.29279%
NumberCarsO...	String	17.70721%	27.6798%	21.13314%	9.28676%	30.78684%
DateFirstPurch...	String	17.70721%	41.1464%	79.72026%	42.39459%	81.85672%

Dataset-level quality

Model	Quality
Gen. intensity	31.15839%
Granularity	41.96925%
N.-U. entropy	28.29649%
Discernibility	82.26754%
Average class size	99.97356%
Record-level squared error	32.42591%
Attribute-level squared error	81.75896%

Figure 27: Data Quality Model Output Data – K-anonymity

The data quality results for the attribute level quality models are shown in Table 12 below.

The following attribute level data quality measures were implemented, namely:

1. Gen Intensity
2. Granularity
3. N-U. entropy
4. Squared Error

Table 12: Attribute Level Data Quality Results

Attribute	Gen Intensity (%)	Granularity (%)	N-U. entropy (%)	Squared error (%)
BirthDate	41.146	79.459	32.742	81.658
MaritalStatus	0	0	0	0
Gender	0	0	0	0
YearlyIncome	49.724	38.986	10.255	64.822
TotalChildren	41.146	49.375	29.610	67.269
NumberChildren	28.448	34.137	17.615	54.218
EnglishEducation	0	0	0	0
HouseOwnerFlag	82.292	82.292	81.085	82.292
NumberCarsOwned	27.679	29.133	9.286	30.786
DateFirstPurchased	41.146	79.720	42.394	81.856

Furthermore, data set level quality models were also tested. In Table 13 below, the data quality results for the seven data set level quality models are shown. These quality models are:

- Gen intensity
- Granularity
- N U entropy
- Discernibility
- Average class size
- Record-level squared error
- Attribute-level squared error

Table 13: Data set Level Data Quality Results

Model	Quality
Gen intensity	35.158
Granularity	41.969
N U entropy	28.296
Discernibility	82.267
Average class size	99.973
Record-level squared error	32.425
Attribute-level squared error	81.758

6.2.3 L-diversity Data Utility Results

In this section the data quality obtained for the output data for the l-diversity privacy model are shown. As shown for k-anonymity previously, the attribute-level as well as the data set-level quality results are again displayed for the l-diversity output data.

Data Quality Models

In Figure 28 below, the results of the testing for data quality when using the l-diversity privacy model is shown.

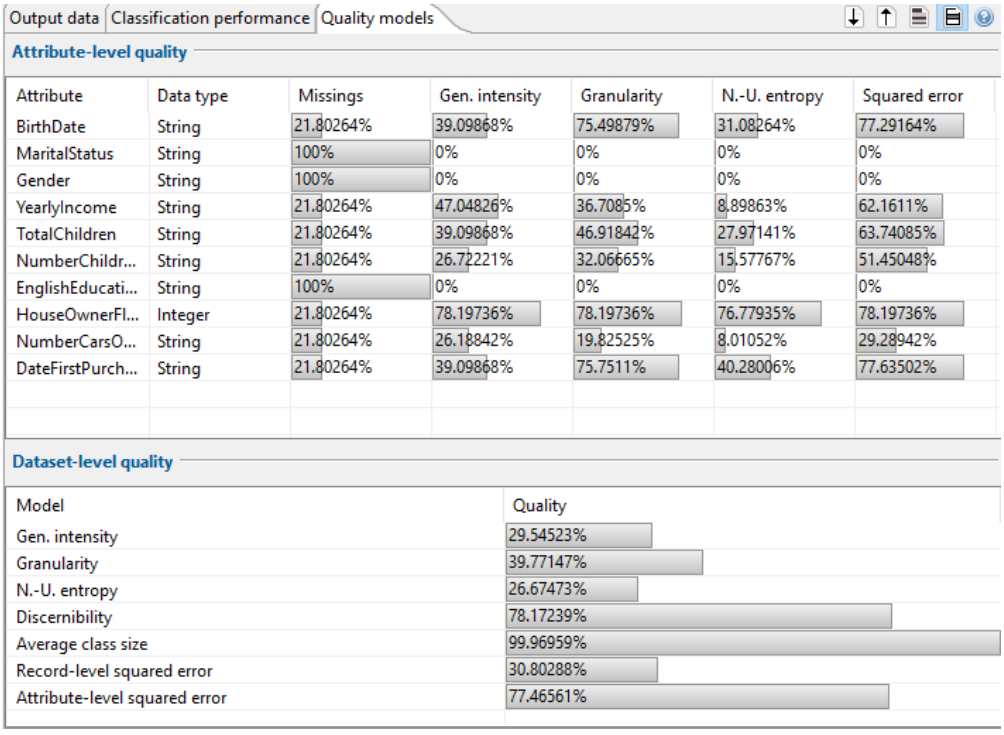


Figure 28: Data Quality Model Output Data –L-diversity

The data quality results for the four attribute-level quality models are shown in Table 14 below. All four attribute level quality models are shown below.

Table 14: Attribute Level Data Quality Results (L-diversity)

Attribute	Gen Intensity(%)	Granularity (%)	N-U. entropy (%)	Squared error (%)
BirthDate	39.098	75.498	31.082	77.291
MaritalStatus	0	0	0	0
Gender	0	0	0	0
YearlyIncome	47.048	36.708	8.898	62.161
TotalChildren	39.098	46.918	27.971	63.740
NumberChildren	27.722	32.066	15.577	51.450
EnglishEducation	0	0	0	0
HouseOwnerFlag	78.197	78.197	76.779	78.197
NumberCarsOwned	26.188	19.825	8.010	29.289
DateFirstPurchased	39.098	75.751	40.280	77.635

Furthermore, the data quality results for the seven data set level quality models are shown in Table 15 below.

Table 15: Data set Level Data Quality Results (L-diversity)

Model	Quality (%)
Gen intensity	29.545
Granularity	39.771
N U entropy	26.674
Discernibility	78.172
Average class size	99.969
Record-level squared error	30.802
Attribute-level squared error	77.465

A further discussion on the analysis of the data quality models that were tested is presented in the next chapter.

6.3 Analysing Privacy Risks

In this section, the privacy risks post-anonymisation of the data set are analysed. The re-identification risks are measured with the results shown below.

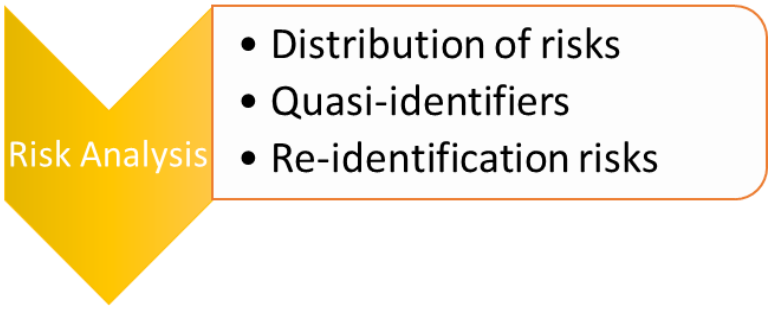


Figure 29: Privacy Risks Considerations

6.3.1 Privacy Risks Testing

When analysing privacy risks, the following features can be analysed:

1. Within the records in the data set, the distribution of re-identification risks can be analysed;
2. A combination of quasi-identifiers can be analysed with regards to re-identification risks;
3. Overview of re-identification risks measures.

The key risk analysis measure that pertains to this study is the third option listed above, namely re-identification risks measures. This functionality relates directly to the testing of the hypothesis and will be used in this study.

6.3.1.1 Re-identification Risks

When considering the re-identification risks of the output data set, several measures for re-identification risks are proposed. When estimating privacy risks for an output data set, attacker models can be used. When measuring privacy risk, three different common attacker models are available [75]:

Prosecutor Attacker Model: it is assumed that the attacker has known knowledge that an individual is already contained in the data set;
Journalist Attacker Model: it is assumed that an attacker has no information about an individual in the data set or their background data;
Marketer Attacker Model: the assumption for this attacker model is that the attacker attacks a large number of individuals contained in a data set and is not particularly interested in de-identifying a single individual record.

Below are the key measures that were used when determining the privacy risks associated for the data set in this study. These key measures for re-identification risks are detailed below:

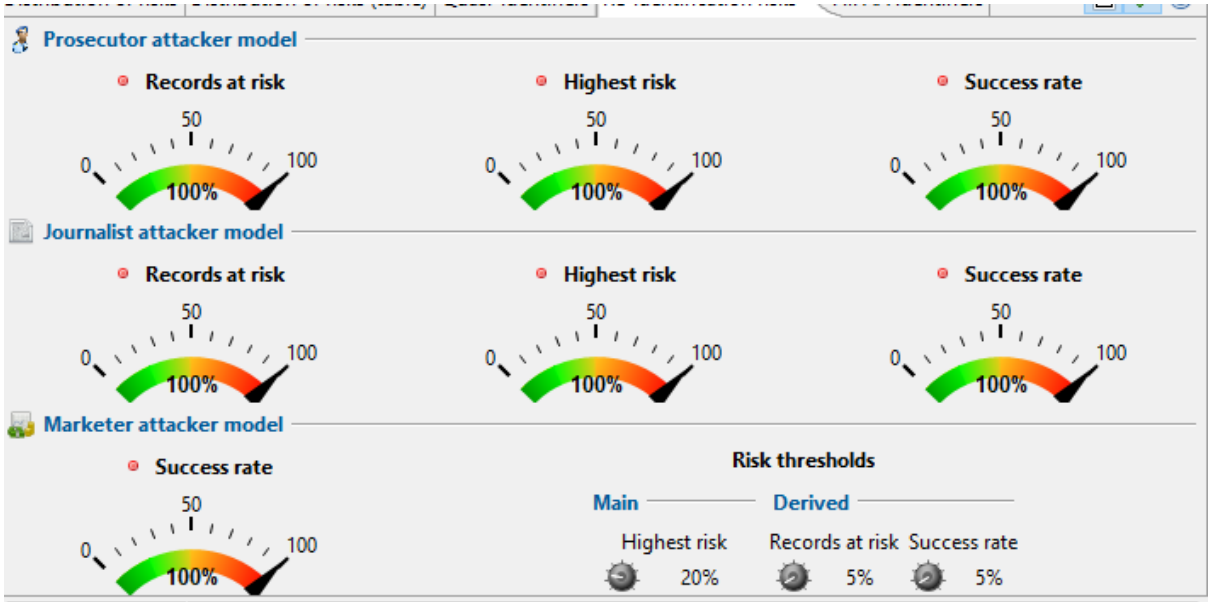
Lowest Prosecutor Risk: this measure shows (as a percentage) the lowest risk of re-identification when considering the prosecutor attacker model.
Records affected by lowest risk: this measure shows the percentage of records (from the total record set) that are affected by the lowest risk when using the prosecutor attacker model.
Average prosecutor risk: this measure shows (as a percentage) the average risk of re-identification when considering the prosecutor attacker model.
Highest prosecutor risk: this measure shows (as a percentage) the highest risk of re-identification when using the prosecutor attacker model.
Records affected by highest risk: this measure shows the percentage of records (from the total record set) that are affected by the highest risk when using the prosecutor attacker model
Estimated prosecutor risk: this measure shows (as a percentage) the estimated risk of re-identification when considering the prosecutor attacker model.
Estimated journalist risk: this measure shows (as a percentage) the estimated risk of re-identification when considering the journalist attacker model.
Estimated marketer risk: this measure shows (as a percentage) the estimated risk of re-identification when considering the marketer attacker model.

6.3.2 K-anonymity Privacy Risks Results

In this section the privacy risk results for the k-anonymity privacy model are shown.

6.3.2.1 Re-identification Risks (K-anonymity)

In Figure 30 below, the results of the testing for privacy of the input data set before anonymisation is performed are shown. Each of the re-identification risks are measured as a percentage before the anonymisation is performed.

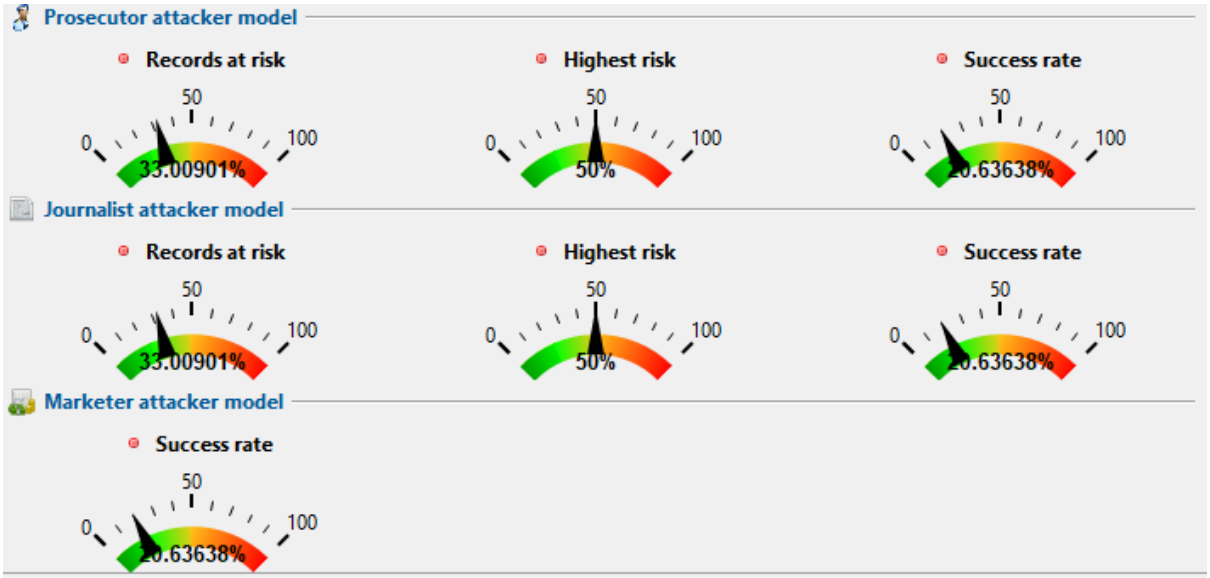


Re-identification risks | Population uniques | Quasi-identifiers

Measure	Value [%]
Lowest prosecutor risk	100%
Records affected by lowest risk	100%
Average prosecutor risk	100%
Highest prosecutor risk	100%
Records affected by highest risk	100%
Estimated prosecutor risk	100%
Estimated journalist risk	100%
Estimated marketer risk	100%
Sample uniques	100%
Population uniques	100%
Population model	ZAYATZ
Quasi-identifiers	BirthDate, DateFirstPurchase, EnglishEducation, Gender, House...

Figure 30: Re-identification Risks of Input Data set (K-anonymity)

In Figure 31 below, the results of the testing for privacy risks when using the k-anonymity privacy model are shown. Each of the re-identification risks are measured as a percentage after the anonymisation was performed.



Measure	Value [%]
Lowest prosecutor risk	5.55556%
Records affected by lowest risk	0.23667%
Average prosecutor risk	20.63638%
Highest prosecutor risk	50%
Records affected by highest risk	10.78167%
Estimated prosecutor risk	50%
Estimated journalist risk	50%
Estimated marketer risk	20.63638%
Sample uniques	0%
Population uniques	0%
Population model	DANKAR
Quasi-identifiers	BirthDate, DateFirstPurchase, EnglishEducation, Gender, House...

Figure 31: Re-identification Risks of Output Data set (K-anonymity)

A summary of the re-identification risks, both before and after anonymisation, is shown side-by-side in Table 16 below. Further analysis into these results, and the resultant comparison table, will be considered in the following chapter on Analysis of Results.

Table 16: K-anonymity Privacy Risks Results (Before and After Anonymisation)

Measure	K-Anonymity Before Anonymisation (%)	K-Anonymity After Anonymisation (%)
Lowest Prosecutor Risk	100	5.55
Records affected by lowest risk	100	0.23
Average prosecutor risk	100	20.63
Highest prosecutor risk	100	50
Records affected by highest risk	100	10.78
Estimated prosecutor risk	100	50
Estimated journalist risk	100	50
Estimated marketer risk	100	20.63
Prosecutor Risk success rate	100	20.63
Journalist Risk success rate	100	20.63
Marketer Risk success rate	100	20.63

6.3.3 L-diversity Privacy Risks Results

In this section the privacy risk results for the l-diversity privacy model are shown.

6.3.3.1 Re-identification Risks (L-diversity)

In Figure 32 below, the results of the testing for privacy of the input data set before anonymisation is performed are shown. Each of the re-identification risks are measured as a percentage before l-diversity anonymisation is done.

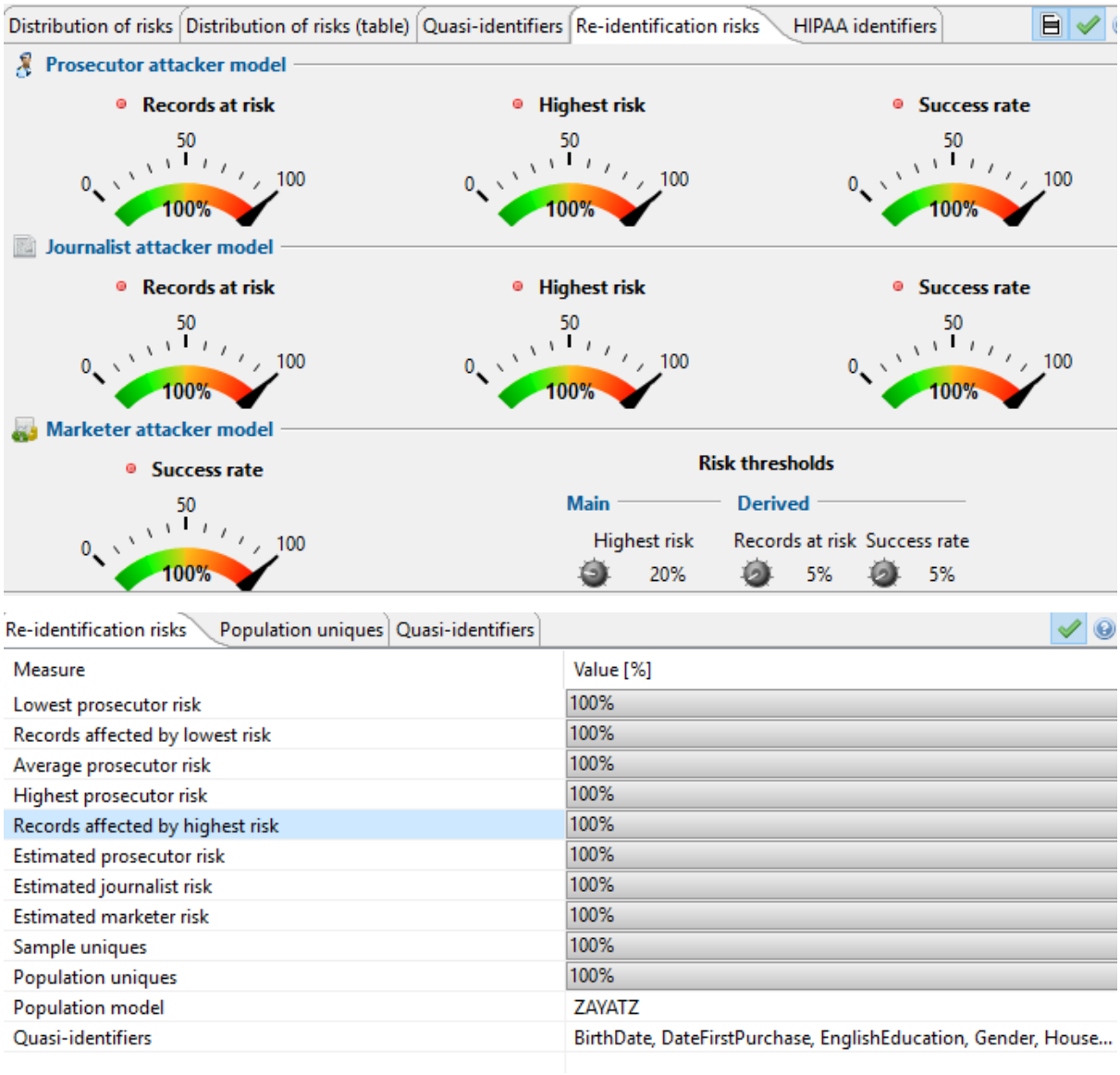
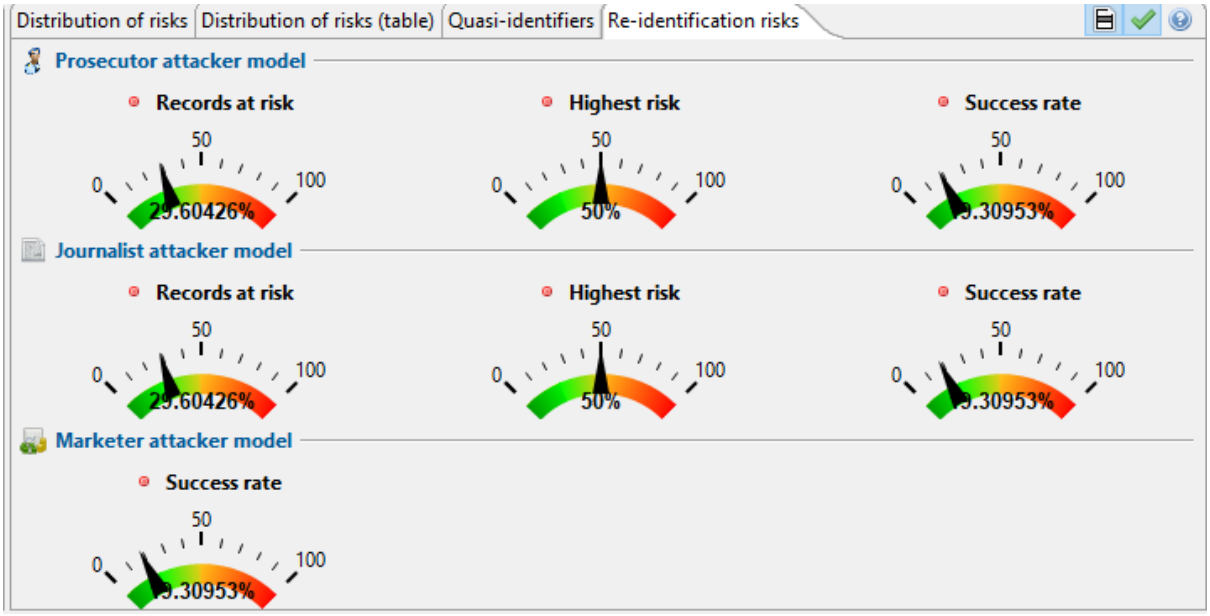


Figure 32: Re-identification Risks of Input Data set (L-diversity)

The resultant Figure 33 below shows the results of the testing for privacy risks after implementing the l-diversity privacy model. Each of the re-identification risks are measured as a percentage after the anonymisation was performed.



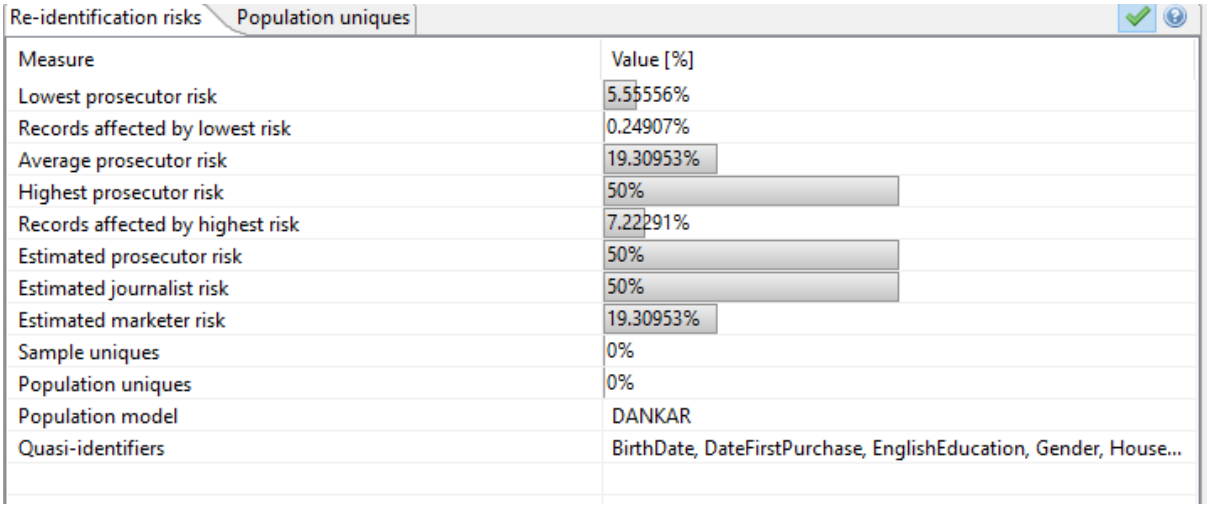


Figure 33: Re-identification Risks of Output Data set (L-diversity)

In Table 17 below a summary of the re-identification risk and individual risk measures, both before and after anonymisation using the l-diversity privacy model, are shown side-by-side.

Table 17: L-diversity Privacy Risks Results (Before and After Anonymisation)

Measure	L-Diversity Before Anonymisation (%)	L-Diversity After Anonymisation (%)
Lowest Prosecutor Risk	100	5.55
Records affected by lowest risk	100	0.24

Average prosecutor risk	100	19.30
Highest prosecutor risk	100	50
Records affected by highest prosecutor risk	100	7.22
Estimated prosecutor risk	100	50
Estimated journalist risk	100	50
Estimated marketer risk	100	19.30
Prosecutor Risk success rate	100	19.30
Journalist Risk success rate	100	19.30
Marketer Risk success rate	100	19.30

A discussion on the significance of the privacy risk measures post-anonymisation is done in the next chapter.

6.4 Summary

In this chapter, the results from the design of the proof of concept and implementation thereof are documented. The various data utility measures, for both attribute level and data set level measures, were reported. Furthermore, a side-by-side comparison of three privacy risk attacker models, namely prosecutor risk attacker model, marketer risk attacker model, and journalist risk attacker model were also presented. In the next chapter, an analysis of the results detailed in this chapter are presented.

7 ANALYSIS OF RESULTS

In the previous chapter, the results as observed from the testing were noted and documented. The aim of this chapter is to show the clear results from the application of the various privacy models, data utility models, and risk attacker models to the data set. This chapter will also show what the results imply and whether the hypothesis and research succeeded or not.

7.1 Interpretation of Findings

In this section, the results for the overall transformations, data utility, and privacy risk results are analysed.

7.1.1 Solution Space Results Analysis

In the solution space, the transformations that were applied to the data set were analysed. These transformations were then captured in Table 18 below.

Table 18: Combined Solution Space Transformations

Privacy Model	Solution Space Transformations
K-anonymity	13,864
L-diversity	13,864

Table 18 above combines the results from both the k-anonymity and l-diversity privacy models when referring to the number of transformations as a result of the anonymisation. It can be clearly seen that for both the k-anonymity and l-diversity privacy models, the number of transformations in the solution space as a result is the same.

7.1.2 Data Utility Results Analysis

In the previous chapter the results for the data quality were recorded. This section will discuss the results of those data quality outputs in detail and in relation to the aims of this research.

7.1.2.1 Attribute Level Data Quality

The following four attribute level data quality measures were implemented.

1. Gen Intensity
2. Granularity
3. N-U. entropy
4. Squared Error

Data quality was measured using the quasi-identifiers as configured previously.

Table 19 represents (in tabular form) the comparison of the gen intensity attribute level quality models for k-anonymity and l-diversity. A side-by-side comparison of both privacy models is shown for all the quasi-identifiers in the data set.

Table 19: Attribute Level Data Quality – Gen Intensity

Attribute	**K-Anonymity Gen Intensity** (%)	**L-Diversity Gen Intensity** (%)
BirthDate	41.146	39.098
MaritalStatus	0	0
Gender	0	0
YearlyIncome	49.724	47.048
TotalChildren	41.146	39.098
NumberChildren	28.448	27.722
EnglishEducation	0	0
HouseOwnerFlag	82.292	78.197
NumberCarsOwned	27.679	26.188
DateFirstPurchased	41.146	39.098

The ten quasi-identifier attributes were analysed against each attribute level data quality model. As shown by the results above, for every attribute in the data set the k-anonymity privacy model performed better than the l-diversity model. Therefore, k-anonymity retained a higher level of data quality when the gen intensity data quality model is applied.

Table 20: Attribute Level Data Quality – Granularity

Attribute	**K-Anonymity Granularity** (%)	**L-Diversity Granularity** (%)
BirthDate	79.459	75.498
MaritalStatus	0	0
Gender	0	0
YearlyIncome	38.986	36.708
TotalChildren	49.375	46.918
NumberChildren	34.137	32.066
EnglishEducation	0	0
HouseOwnerFlag	82.292	78.197
NumberCarsOwned	29.133	19.825
DateFirstPurchased	79.720	75.751

Table 20 represents (in tabular form) the comparison of the granularity attribute level quality models for k-anonymity and l-diversity. A similar side-by-side comparison is shown.

As shown by the results above, for every attribute in the data set, the k-anonymity privacy model performed better than the l-diversity model. Therefore, it can be seen that k-anonymity retained a higher level of data quality when the granularity data quality model is applied.

Table 21: Attribute Level Data Quality – N-U. Entropy

Attribute	**K-Anonymity** **N-U. entropy** (%)	**L-Diversity** **N-U. entropy** (%)
BirthDate	32.742	31.082
MaritalStatus	0	0
Gender	0	0
YearlyIncome	10.255	8.898
TotalChildren	29.610	27.971
NumberChildren	17.615	15.577
EnglishEducation	0	0
HouseOwnerFlag	81.085	76.779
NumberCarsOwned	9.286	8.010
DateFirstPurchased	42.394	40.280

Table 21 represents (in tabular form) the comparison of the n-u. entropy attribute-level quality model for k-anonymity and l-diversity. The side-by-side comparison is shown in the figure above.

As shown by the results above, for every attribute in the data set the k-anonymity privacy model performed better than the l-diversity model. Therefore, k-anonymity retained a higher level of data quality when the n-u. entropy data quality model is applied.

Table 22: Attribute Level Data Quality – Squared Error

Attribute	**K-Anonymity** **Squared error** (%)	**L-Diversity** **Squared error** (%)
BirthDate	81.658	77.291
MaritalStatus	0	0
Gender	0	0
YearlyIncome	64.822	62.161
TotalChildren	67.269	63.740
NumberChildren	54.218	51.450
EnglishEducation	0	0

HouseOwnerFlag	82.292	78.197
NumberCarsOwned	30.786	29.289
DateFirstPurchased	81.856	77.635

Table 22 represents (in tabular form) the comparison of the squared error attribute-level quality model for k-anonymity and l-diversity. A side-by-side comparison is shown in the figure above with the k-anonymity performing better than l-diversity for the squared error data quality model.

In summary, as shown by the results, for every attribute in the data set for the gen intensity, granularity, N-U entropy and squared error data quality models, the k-anonymity privacy model performed better than the l-diversity model of all four measures. Therefore, it can be seen that k-anonymity retained a greater level of data quality for all attribute level data quality measures after the anonymisation process.

7.1.2.2 Data set Level Data Quality

For the data set level data quality models, quality was measured for the entire set of quasi-identifiers and not individual quasi-identifiers as was done with attribute level data quality models.

Table 23: Combined Data set Level Quality Results

Quality Model	**K-Anonymity (%)**	**L-Diversity (%)**
Gen intensity	35.158	29.545
Granularity	41.969	39.771
N U entropy	28.296	26.674
Discernibility	82.267	78.172
Average class size	99.973	99.969
Record-level squared error	32.425	30.802
Attribute-level squared error	81.758	77.465

Table 23 above represents (in tabular form) the comparison of the k-anonymity and l-diversity results. This is when the full data set of quasi-identifiers is referenced. The results are shown for the following seven data set level quality models:

1. Gen intensity
2. Granularity
3. N U entropy
4. Discernibility
5. Average class size
6. Record-level squared error
7. Attribute-level squared error

A side-by-side comparison is shown in the figure above. As shown by the results, for all seven data set level quality models the k-anonymity privacy model performs better than the l-diversity model obtaining a higher percentage of quality. It is worthwhile to note that the k-anonymity privacy model retained a higher level of data quality than l-diversity for all data set level quality measures.

7.1.3 Privacy Risks Results Analysis

In the previous chapter, the privacy risks for each privacy model were discussed and recorded. Furthermore, this section discusses the results of those outputs in more detail relating back to the aim of this study.

7.1.3.1 Privacy Attacker Models

The following three privacy risk attacker models were analysed, namely:

1. Prosecutor Attack Model
2. Journalist Attack Model
3. Marketer Attack Model

7.1.3.2 Privacy Risk Measures

The following privacy risk measures were analysed for both k-anonymity and l-diversity privacy models. An overview of each privacy risk measure is provided below:

Lowest Prosecutor Risk: this measure shows (as a percentage) the lowest risk of re-identification when considering the prosecutor attacker model.

Records affected by lowest risk: this measure shows the percentage of records (from the total record set) that are affected by the lowest risk when using the prosecutor attacker model.

Average prosecutor risk: this measure shows (as a percentage) the average risk of re-identification when considering the prosecutor attacker model.

Highest prosecutor risk: this measure shows (as a percentage) the highest risk of re-identification when considering the prosecutor attacker model.

Records affected by highest risk: this measure shows the percentage of records (from the total record set) that are affected by the highest risk when using the prosecutor attacker model.

Estimated prosecutor risk: this measure shows (as a percentage) the estimated risk of re-identification when considering the prosecutor attacker model.

Estimated journalist risk: this measure shows (as a percentage) the estimated risk of re-identification when considering the journalist attacker model.

Estimated marketer risk: this measure shows (as a percentage) the estimated risk of re-identification when considering the marketer attacker model.

7.1.3.3 Privacy Risks Results

Re-identification risk outputs for the prosecutor, marketer, and journalist attacker models and their various measures are shown below. These measures are listed for the l-diversity and k-anonymity privacy model next to each other for comparison. Please refer to Table 24.

Table 24: Combined Re-identification Risks Results

Measure	K-Anonymity (%)	L-Diversity (%)
Lowest Prosecutor Risk	5.55	5.55
Records affected by lowest risk	0.23	0.24
Average prosecutor risk	20.63	19.30
Highest prosecutor risk	50	50
Records affected by highest risk	10.78	7.22
Estimated prosecutor risk	50	50
Estimated journalist risk	50	50
Estimated marketer risk	20.63	19.30
Prosecutor Risk success rate	20.63	19.30
Journalist Risk success rate	20.63	19.30
Marketer Risk success rate	20.63	19.30

Table 24 shows the key measures that were used when determining the privacy risks associated for the data set in this study. It can be noted, if an outcome of the re-identification is high, the privacy model being used does not protect individual identifying information adequately in the data set.

Finally, the prosecutor risk success rate, journalist risk success rate, and marketer risk success rate (highlighted in bold above) are the most important measures to consider in this study. These measures show the overall success rate of each attacker model when using relevant privacy models, and, therefore, the analysis of these measures refer directly to the aims of the research and hypothesis.

As shown by the results in Table 24, the k-anonymity privacy model has a marginally higher rate of success when referring to all three attacker models. It is 1.36% higher than l-diversity for all three attacker models.

7.2 Evaluation of Results

7.2.1 Data Utility Evaluation

In summary of the above, when referring to the attribute level data quality evaluation for each of the four attribute level data quality models, the k-anonymity privacy model performed better than the l-diversity model in every instance. Therefore, k-anonymity retained a higher level of data quality than l-diversity when measuring data set level quality.

When evaluating the outputs against the seven data set level quality models, for every model the k-anonymity privacy model performed better than the l-diversity privacy model.

It is consistent in this evaluation that for both attribute level and data set level quality models, k-anonymity retained the higher level of quality than l-diversity.

7.2.2 Privacy Risks Evaluation

When measuring re-identification risks, it was clearly seen that the l-diversity privacy model performed marginally better than k-anonymity. L-diversity had a 1.36% lower chance of re-identification than k-anonymity. Therefore, when evaluating the customer data set for privacy risk measure, the l-diversity privacy model is the better model between the two.

7.3 Summary

In this chapter, the analysis of the results presented in the previous chapter is documented. Both data quality models and privacy risk models were evaluated, and the outcomes captured. From a data quality perspective, k-anonymity results achieve a higher data quality output than l-diversity. However, when considering privacy risks, the l-diversity privacy model resulted in a lower re-identification risk across all three attacker models. In the next chapter, the results of the analysis are summarised to obtain a conclusion to the study, relating back to the aim as initially proposed.

8 CONCLUSION

8.1 Research Summary

In the previous chapter the analysis of the results from the implementation of the two privacy models are shown. In this chapter a summary of the research findings is discussed. As discussed previously, the need to protect personally identifying information is a relatively new concept in South Africa. With the promulgation of the POPI Act, the need for protecting customer data, using a practical approach, is becoming more and more important and necessary. The challenge faced with the implementation of a tool for anonymising data is that tools are not readily available. To this end, this study shows the implementation of two well-known privacy models in a practical, usable way in which to determine which of the anonymisation privacy models is more suited to ensure privacy protection for individuals to ensure POPI compliance. This chapter will summarise the research outcomes in relation to the research aim, research questions as well as research objectives. Finally, the results achieved are presented together with recommendations for future work.

8.2 Discussion

At the beginning of the thesis, research questions were developed in order to meet the aim and objectives of the research. Firstly, to re-iterate, the following hypothesis was formulated:

> Practically implementing the k-anonymity privacy model to anonymise static customer data offers an effective solution to ensure POPI compliance by retaining the highest level of privacy without compromising the utility of the data.

Thereafter, research questions were developed to concisely address the hypothesis and objectives. Table 1 mapped the research questions in relation to the research objectives to clearly show what the study was aiming to achieve. A discussion of the research questions and research objectives follows to show the link back to the hypothesis.

MAIN: Which privacy-preserving anonymisation technique is appropriate for anonymising static customer data for POPI act compliance?
SUB 1: Which privacy-preserving data anonymisation technique is appropriate for practically anonymising static customer data for POPI Act compliance by ensuring the lowest level of privacy risk?
SUB 2: Can privacy-preserving data anonymisation techniques be practically applied to anonymise static customer data for POPI compliance in a way that preserves the greatest data utility?
SUB 3: What are the effects of the levels of privacy risk and data utility in ensuring POPI Act compliance of static customer data?

MAIN: Which privacy-preserving anonymisation technique is appropriate for anonymising static customer data for POPI act compliance?

- **Research Objective 1:** Identify and evaluate existing data anonymisation techniques
- **Research Objective 2:** Identify a data anonymisation tool to practically perform the data anonymisation process
- **Research Objective 3:** Implement two privacy models within a data anonymisation tool

In order to address the main research question and research objectives above, an appropriate anonymisation tool was selected so that the anonymisation techniques as provided in the literature review could be tested. The implementation of an anonymisation tool was done in the environment as outlined in Chapter 4. Static customer data was sourced and used to simulate real-world customer banking data. Two privacy models were selected and configured for evaluation, namely k-anonymity and l-diversity.
The results showed both privacy models evaluated can be adequately applied to anonymise static customer data in such a way to ensure adherence to the rules of each of the privacy models.

SUB 1: Which privacy-preserving data anonymisation technique is appropriate for practically anonymising static customer data for POPI Act compliance by ensuring the lowest level of privacy risk?

- **Research Objective 4:** Compare and contrast the results of the privacy models with one another with regards to the lowest level of privacy risk to ensure POPI Act adherence

When taking privacy risks and data utility into consideration, l-diversity performed better than k-anonymity for all re-identification risks that were tested. Considering that protecting the privacy of an individual's information is significantly more important that data utility preservation, l-diversity outperformed k-anonymity by 1.36%.

SUB 2: Can privacy-preserving data anonymisation techniques be practically applied to anonymise static customer data for POPI compliance in a way that preserves the greatest data utility?

- **Research Objective 5:** Compare and contrast the results of the privacy models with one another with regards to the highest level of utility retained so that the information is still useful

As anonymisation of personally identifiable information can result in data that is not useful; the line between the level of data utility and the risk of re-identification is very fine. Therefore, this research aimed to quantify these output measures to allow for a fair comparison of the anonymisation privacy models selected for POPI compliance. In order to do this, data quality models were applied to the data set to measure the level of data quality achieved for each of the two privacy models.

For attribute-level data quality measures of precision, granularity, non-uniform entropy and squared error, k-anonymity results were higher than l-diversity.

For data set level data quality models, namely average class size, discernibility, ambiguity and record-level squared error k-anonymity once again performed better than l-diversity. In summary, k-anonymity performed better than l-diversity on all data quality models and measures.

When referring to privacy risks, the results were closer than expected. The following three attacker models were analysed, namely prosecutor attack model, journalist attack model, and marketer attack model. As shown by the results, the k-anonymity privacy model had a marginally higher rate of success when looking at the results of all three attacker models success rates. K-anonymity results were 1.36% higher for re-identification risk than l-diversity results for all three attacker models. Therefore, l-diversity offers a lower chance of re-identification than k-anonymity.

SUB 3: What are the effects of the levels of privacy risk and data utility in ensuring POPI Act compliance of static customer data?

- **Research Objective 6:** Establish the effects of privacy risks and data utility for static customer data in relating to POPI Act compliance

Finally, in showing the results, it was split evenly between the k-anonymity and l-diversity privacy model with regards to the two measures tested, namely data utility and privacy risks. To summarise, k-anonymity retained the most data utility whilst l-diversity offered the lowest privacy risk. Therefore, seeing that data privacy preservation is the key measure over data utility when taking personally identifiable information into consideration, anonymisation techniques that are implemented by the l-diversity privacy model are more successful at protecting privacy.

8.3 Results Achieved

In summary, the results of this study and the objectives were accomplished. In reviewing the hypothesis of this study, the tool selected successfully anonymised the static customer data set that was provided as input. The software platform allowed for the configuration of the anonymisation for each of the two data privacy models, namely l-diversity and k-anonymity, taking into consideration the generalisation of the individual attributes where appropriate.
The outcomes, once the data anonymisation process was concluded, presented the resultant measures for data utility and privacy risks for re-identification.

In the first set of results, the data utility in relation to data quality models applied were analysed. The clear outcome was that on both attribute level data quality and data set level data quality, k-anonymity was better at retaining data utility than l-diversity.
When the privacy risk attacker models were applied to the data set for each of the three attacker models, l-diversity showed a lower risk of re-identification than k-anonymity.

When taking both utility and privacy into consideration, neither privacy model selected is the clear winner. The k-anonymity privacy model is a better choice for data utility, and l-diversity privacy model is a better choice for privacy preservation. However, when referring to the hypothesis that the k-anonymity privacy model is more suitable to anonymise data with regards to the POPI Act, an outcome is that the l-diversity model is the more successful model of the two. Therefore, the hypothesis that k-anonymity is more suitable for data anonymisation for POPI compliance is not true when considering the measure of privacy risk. L-diversity is the preferred privacy model.

Finally, considering that the POPI Act is still awaiting the final step to be promulgated into law, financial institutions and specifically banking institutions will have the necessary time to conduct further experiments to practically implement and apply data anonymisation techniques in their day-to-data processing of data and information. This is because there will be a period until the law is made effective. If the key aspect of ensuring that the right to privacy of an individual is always guaranteed, the analysis of customer banking data by external or third parties can be enforced.

8.4 Recommendations for Future Research

With the promulgation of the POPI Act still to come, and with data privacy implications for companies cutting across jurisdictions and countries, ensuring data privacy for individuals will become a common place in the future in South Africa. Banking institutions will need to be more agile to keep up with ensuring regulatory compliance in relation to their competitors.

When using the tool ARX for anonymisation of data, the alternative for a public Application Programming Interface (API) instead of using the graphical interface is provided. All features and functionality are available within the API. This would allow for the de-identification methods to be made available to other software systems for scenarios where this would be easier.
Therefore, a key recommendation for future research would be to implement the anonymisation tool ARX in an API environment for integration with a Java platform using the POPI Act as the area of legislation. By doing so, this will allow for the use of ARX for non-static data in an environment.

Seeing that ARX is open-source and not a commercial tool, the development of an option for evaluating other privacy risks in ARX, similar to the HIPAA tab, would be useful. The HIPAA tab refers to the regulations that govern the adherence to the HIPAA Act as outlined in Chapter 2. Eighteen key identifiers have been specified for modification or removal by the Safe Harbour method of HIPAA. A similar implementation for modifying or removing known identifiers in ARX could automatically allow for the detection of POPI identifiers and appropriately anonymise that data. This would be a very useful feature in future.

Once the Information Regulator has proposed clear rules on how customer information can and cannot be processed with regards to the POPI Act, these rules can be written into ARX and applied to customer data in South Africa similar to how the HIPAA tab is structured. As a result, the privacy risk in relation to the POPI Act can then be viewed and analysed by users without any prior knowledge of data anonymisation.

9 REFERENCES

[1] T. Li and N. Li, “On the tradeoff between privacy and utility in data publishing,” in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’09*, 2009, p. 517.

[2] Government of South African, “Protection of Personal Information Act, 2013 Ensuring protection of your personal information and effective access to information.”

[3] A. Cavoukian and K. El Emam, “De-identification Protocols: Essential for Protecting Privacy Information and Privacy Commissioner Ontario, Canada,” 2014.

[4] HHS.gov, “Health Insurance Portability and Accountability Act,” 2016. [Online]. Available: https://www.hhs.gov/hipaa/index.html. [Accessed: 27-Jun-2019].

[5] F. Prasser, F. Kohlmayer, R. Lautenschläger, and K. A. Kuhn, “ARX - A Comprehensive Tool for Anonymizing Biomedical Data,” *AMIA Annu. Symp. Proc.*, pp. 984–993, 2014.

[6] A. Gkoulalas-Divanis, G. Loukides, and J. Sun, “Publishing data from electronic health records while preserving privacy: A survey of algorithms,” 2014.

[7] United Nations Conference on Trade and Development (UNCTAD), “Data protection regulations and international data flows: Implications for trade and development,” *United Nations Publ.*, 2016.

[8] N. Baloyi and P. Kotze, “Are organisations in South Africa ready to comply with personal data protection or privacy legislation and regulations?,” in *2017 IST-Africa Week Conference (IST-Africa)*, 2017, pp. 1–11.

[9] E. and Young, “What happens if we violate PoPI? What is PoPI?,” 2013.

[10] E. Mccallister, K. Scarfone, and T. Grance, “Guide to Protecting the Confidentiality of Personally Identifiable Information (PII) Recommendations of the National Institute of Standards and Technology,” *NIST Spec. Publ. 800-122*, p. 59, 2010.

[11] L. Sweeney, “Achieving k-anonymity privacy protection using generalization and suppression,” 2002.

[12] F. Kohlmayer, F. Prasser, and K. A. Kuhn, “The cost of quality: Implementing generalization and suppression for anonymizing biomedical data with minimal information loss,” *J. Biomed. Inform.*, vol. 58, pp. 37–48, Dec. 2015.

[13] W. D. Eggers, R. Hamill, A. Ali, and J. Hersey, “Data as the new currency Government’s role in facilitating the exchange,” 2013.

[14] R. Mendes and J. P. Vilela, “Privacy-Preserving Data Mining: Methods, Metrics, and Applications,” *IEEE Access*, vol. 5, pp. 10562–10582, 2017.

[15] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, “State-of-the-art in privacy preserving data mining,” *ACM SIGMOD Rec.*, vol. 33, no. 1, p. 50, Mar. 2004.

[16] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, “Privacy-preserving data publishing,” *ACM Comput. Surv.*, vol. 42, no. 4, pp. 1–53, Jun. 2010.

[17] B. Fung, K. Wang, A. Fu, and S. Philip, *Introduction to privacy-preserving data publishing: Concepts and techniques*. 2010.

[18] K. Benitez and B. Malin, “Evaluating re-identification risks with respect to the HIPAA privacy rule.”

[19] K. El Emam *et al.*, “A Globally Optimal k-Anonymity Method for the De-Identification of Health Data,” *J. Am. Med. Informatics Assoc.*, vol. 16, no. 5, pp. 670–682, 2009.

[20] O. Tomashchuk, D. Van Landuyt, D. Pletea, K. Wuyts, and W. Joosen, “A Data Utility-Driven Benchmark for De-identification Methods,” 2019, pp. 63–77.

[21] L. Bolognini and C. Bistolfi, “Pseudonymization and impacts of Big (personal/anonymous) Data processing in the transition from the Directive 95/46/EC to the new EU General Data Protection Regulation,” 2017.

[22] South African DEPARTMENT OF JUSTICE AND CONSTITUTIONAL DEVELOPMENT, “POPI Act, 2013. Regulations relating to the protection of personal information,” 2018.

[23] J. Botha, M. M. Grobler, J. Hahn, and M. Eloff, “A High-Level Comparison Between the South African Protection of Personal Information Act and International Data Protection Laws,” *Int. Conf. Cyber Warf. Secur.*, no. March, p. 57, 2017.

[24] M. De Bruyn, “The Protection of Personal Information Act (POPI) - Impact on South Africa,” *Int. Bus. Econ. Res. J.*, 2014.

[25] South African Institute of Chartered Accountants, “Information Regulator POPI Act No 4 of 2013,” 2016.

[26] Michaelsons, “POPI Act - Protection of Personal Information,” 2014. [Online]. Available: https://www.michalsons.com/focus-areas/privacy-and-data-

protection/popi-act-protection-of-personal-information. [Accessed: 10-Jan-2020].

[27] Wired UK, “What is GDPR? The summary guide to GDPR compliance in the UK | WIRED UK.” [Online]. Available: https://www.wired.co.uk/article/what-is-gdpr-uk-eu-legislation-compliance-summary-fines-2018. [Accessed: 10-Jan-2020].

[28] UK Government, “GDPR Data Protection Act 2018,” 2018.

[29] M. Katurura and L. Cilliers, “The extent to which the POPI act makes provision for patient privacy in mobile personal health record systems,” *2016 IST-Africa Conf. IST-Africa 2016*, pp. 1–8, 2016.

[30] C. Tikkinen-Piri, A. Rohunen, and J. Markkula, “EU General Data Protection Regulation: Changes and implications for personal data collecting companies,” *Comput. Law Secur. Rev.*, vol. 34, no. 1, pp. 134–153, 2018.

[31] GDPR.EU, “What is GDPR, the EU’s new data protection law? - GDPR.eu.” [Online]. Available: https://gdpr.eu/what-is-gdpr/. [Accessed: 10-Jan-2020].

[32] I. Jolly, “Data protection in the United States: overview,” *Thomson Reuters: Practical Law*, 2018. [Online]. Available: https://uk.practicallaw.thomsonreuters.com/6-502-0467?transitionType=Default&contextData=(sc.Default)&firstPage=true&bhcp=1. [Accessed: 19-Jan-2019].

[33] D. of H. and H. S. (US), “HIIPA Omnibus Rule,” vol. 78, no. 17, pp. 1–138, 2013.

[34] L. Sweeney, “Simple Demographics Often Identify People Uniquely,” Sweeney, 2000.

[35] “Netflix South Africa.” [Online]. Available: https://www.netflix.com/za/. [Accessed: 27-Nov-2019].

[36] A. Narayanan and V. Shmatikov, “Robust de-anonymization of large sparse datasets,” in *Proceedings - IEEE Symposium on Security and Privacy*, 2008, pp. 111–125.

[37] Harvard Medical School, “The Harvard Personal Genome Project (PGP) – enabling participant-driven science,” *The Harvard Personal Genome Project*, 2005. [Online]. Available: https://pgp.med.harvard.edu/. [Accessed: 21-Jan-2019].

[38] L. Sweeney, A. Abu, and J. Winn, “Identifying Participants in the Personal Genome Project by Name (A Re-identification Experiment),” *SSRN Electron. J.*, pp. 1–4, 2013.

[39] H. Jin, “Practical issues on privacy-preserving health data mining,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2007.

[40] D. Kifer and J. Gehrke, “Injecting utility into anonymized datasets,” in *Proceedings of the 2006 ACM SIGMOD international conference on Management of data - SIGMOD ’06*, 2006, p. 217.

[41] K. El Emam, *Guide to the De-Identification of Personal Health Information*. 2013.

[42] F. P. Helmut Spengler, “Protecting Biomedical Data Against Attribute Disclosure,” in *German Medical Data Sciences: Shaping Change – Creative Solutions for Innovative Medicine*, Volume 267., IOS Press Ebooks, 2019, pp. 207–214.

[43] S. Zhong, Z. Yang, and R. N. Wright, “Privacy-enhancing k -anonymization of customer data,” in *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS ’05*, 2005, p. 139.

[44] C. C. Aggarwal and P. S. Yu, “A General Survey of Privacy-Preserving Data Mining Models and Algorithms,” Springer, Boston, MA, 2008, pp. 11–52.

[45] A. Evfimievski, “Randomization in privacy preserving data mining,” *ACM SIGKDD Explor. Newsl.*, vol. 4, no. 2, pp. 43–48, Dec. 2002.

[46] S. P. Reiss, “Practical data-swapping: The first steps,” in *Proceedings - IEEE Symposium on Security and Privacy*, 2012, vol. 9, no. 1, pp. 38–45.

[47] P. Samarati, “Protecting respondents’ identities in microdata release,” *IEEE Trans. Knowl. Data Eng.*, vol. 13, no. 6, pp. 1010–1027, 2001.

[48] M. S. Simi, K. S. Nayaki, and M. S. Elayidom, “An Extensive Study on Data Anonymization Algorithms Based on K-Anonymity,” in *IOP Conference Series: Materials Science and Engineering*, 2017, vol. 225, no. 1, p. 012279.

[49] S. Kisilevich, L. Rokach, Y. Elovici, and B. Shapira, “Efficient multidimensional suppression for k-anonymity,” *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 3, pp. 334–347, Mar. 2010.

[50] Y. Xu, T. Ma, M. Tang, and W. Tian, “A survey of privacy preserving data publishing using generalization and suppression,” *Appl. Math. Inf. Sci.*, vol. 8, no. 3, pp. 1103–1116, 2014.

[51] A. Meyerson and R. Williams, “On the complexity of optimal K-anonymity,” in *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium*

on Principles of database systems - PODS '04, 2004, p. 223.

[52] F. Prasser and F. Kohlmayer, "Putting statistical disclosure control into practice: The ARX data anonymization tool," in *Medical Data Privacy Handbook*, Cham: Springer International Publishing, 2015, pp. 111–148.

[53] T. Li, N. Li, J. Zhang, and I. Molloy, "Slicing: A New Approach for Privacy Preserving Data Publishing," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 3, pp. 561–574, Mar. 2012.

[54] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-Diversity: Privacy Beyond k-Anonymity," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, pp. 3-es, 2007.

[55] J. Eicher, K. A. Kuhn, and F. Prasser, "An experimental comparison of quality models for health data de-identification," *Stud. Health Technol. Inform.*, vol. 245, pp. 704–708, 2017.

[56] L. SWEENEY, "k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY," *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.*, vol. 10, no. 05, pp. 557–570, 2002.

[57] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information," in *Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, 1998, p. 188.

[58] A. Gkoulalas-Divanis and G. Loukides, *Medical data privacy handbook*. 2015.

[59] R. Bild, K. A. Kuhn, and F. Prasser, "SafePub: A Truthful Data Anonymization Algorithm With Strong Privacy Guarantees," *Proc. Priv. Enhancing Technol.*, vol. 2018, no. 1, pp. 67–87, 2018.

[60] NIPO, "GDPR - Discover the Techniques To Protect Personal Data." [Online]. Available: https://www.nipo.com/gdpr-discover-techniques-to-protect-personal-data. [Accessed: 21-Jan-2019].

[61] D. C. Barth-Jones, "The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now," *SSRN Electron. J.*, Jul. 2012.

[62] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and-Diversity."

[63] J. Brickell and V. Shmatikov, "The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing," in *Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD)*, 2008.

[64] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu, “Aggregate query answering on anonymized tables,” in *Proceedings - International Conference on Data Engineering*, 2007, pp. 116–125.

[65] D. Rebollo-Monedero, J. Forné, and J. Domingo-Ferrer, “From t-Closeness-like privacy to postrandomization via information theory,” *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 11, pp. 1623–1636, Nov. 2010.

[66] “ARX - Data Anonymization Tool | A comprehensive software for privacy-preserving microdata publishing.” [Online]. Available: https://arx.deidentifier.org/. [Accessed: 30-Oct-2019].

[67] Electronic Health Information Laboratory, “What de-identification software tools are there? - Electronic Health Information Laboratory.” [Online]. Available: http://www.ehealthinformation.ca/faq/de-identification-software-tools/. [Accessed: 23-Jan-2019].

[68] H. I. Outcomes, “Privacy Analytics’ PARAT 60 The Next Generation De-Identification Software Unlocks The Potential For Faster Better Health Data Analytics,” 2014. [Online]. Available: https://www.healthitoutcomes.com/doc/privacy-analytics-parat-de-identification-software-0001. [Accessed: 23-Jan-2019].

[69] R. Fraser and D. Willison, “Tools for De-Identification of Personal Health Information,” *Canada Heal. Infow.*, no. September, 2009.

[70] S. Netherlands, “μ-ARGUS,” 2018. [Online]. Available: http://research.cbs.nl/casc/mu.htm. [Accessed: 23-Jan-2019].

[71] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, “Incognito,” in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data - SIGMOD ’05*, 2005, p. 49.

[72] University of Texas, “UTD Anonymization ToolBox.” [Online]. Available: http://cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php. [Accessed: 23-Jan-2019].

[73] U. of Texas, “UT DALLAS ANONYMIZATION TOOLBOX MANUAL,” 2012.

[74] M. Templ, “Package ‘sdcMicro’ reference manual,” 2018.

[75] Y. Vural and M. Aydos, “A New Approach to Utility-based Privacy Preserving in Data Publishing.”

[76] R. J. Bayardo and R. Agrawal, “Data privacy through optimal k-anonymization,” in *Proceedings - International Conference on Data Engineering*, 2005, pp. 217–228.

[77] K. Lefevre, D. J. Dewitt, and R. Ramakrishnan, “Multidimensional K-Anonymity,” *Univ. Wisconsin-Madison Dep. Comput. Sci.*, 2005.

[78] V. S. Iyengar and V. S., “Transforming data to satisfy privacy constraints,” in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’02*, 2002, p. 279.

[79] F. Prasser, R. Bild, and K. A. Kuhn, “A Generic method for assessing the quality of De-Identified health data,” *Stud. Health Technol. Inform.*, vol. 228, pp. 312–316, 2017.

[80] C. W. Dawson, *Projects in Computing and Information Systems - A Student’s Guide*, 2nd ed. Pearson Prentice Hall, 2009.

[81] A. Friedman, “Privacy Preserving Data Mining,” Technion - Isreal Institute of Technology, 2011.

[82] “AdventureWorks Sample Database - SQL Server.” [Online]. Available: https://docs.microsoft.com/en-us/sql/samples/adventureworks-install-configure?view=sql-server-ver15. [Accessed: 27-Nov-2019].

[83] P. D. P. Commission, “Guide to basic data anonymisation techniques,” Singapore, 2018.

10 APPENDICES

RESULTS OVERVIEW

10.1 K-anonymity Data Utility Results

The data utility results for the k-anonymity privacy model is shown below. The attribute “YearlyIncome” was selected.

10.1.1 Input Output Data - All Fields

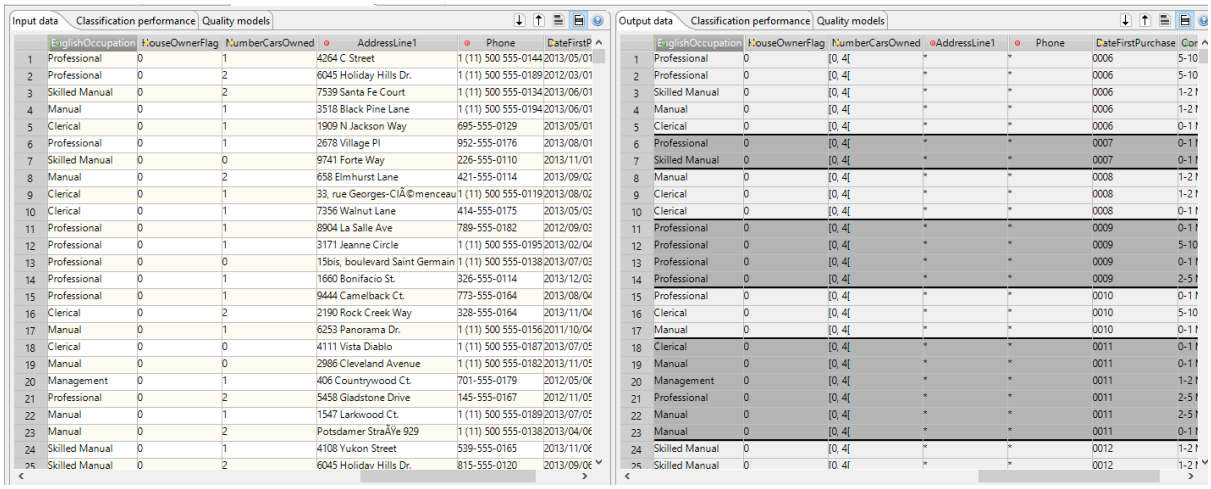


Figure 34: Input and Output Data

The input data set is shown on the left pane and the results after anonymisation is shown on the right pane.

10.1.2 Summary Statistics – Attribute Level

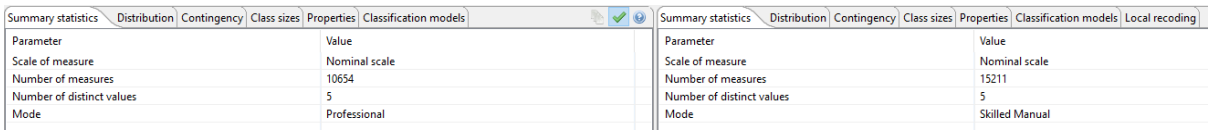


Figure 35: Summary Statistics

The summary statistics for the chosen attribute is shown above.

10.1.3 Frequency Distribution – Attribute level

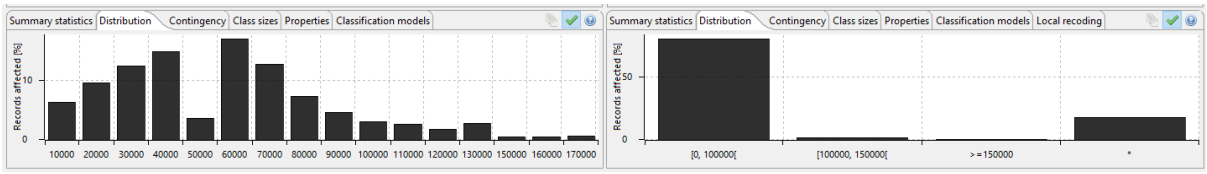


Figure 36: Frequency Distribution

A table or histogram that visualises the frequency distribution of the values of the selected attribute is shown above.

10.1.4 Contingency - Attribute Level across 2 Attributes

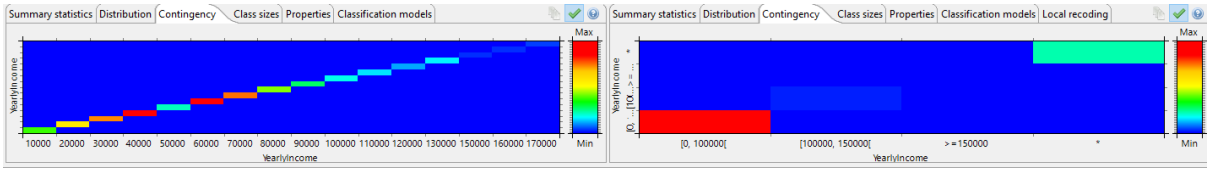


Figure 37: Contingency

The figure above represents a heat map that visually shows the contingency of two selected attributes. Contingency refers to the multivariate frequency distribution of the variables.

10.1.5 Equivalence Classes and Records – All Records

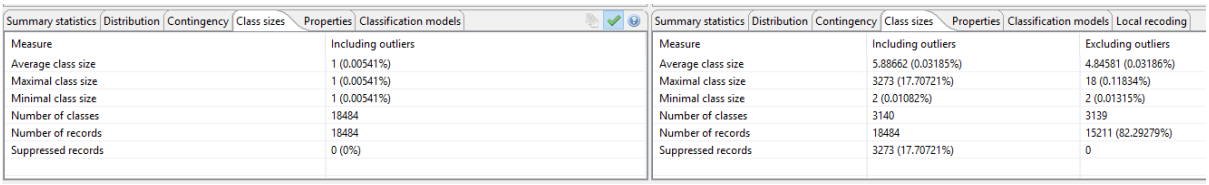


Figure 38: Equivalence Classes and Records

The figure above shows the equivalence classes and records which summarises the information and shows the minimal class size, maximal class size and average size of the equivalence classes.

10.1.6 Properties of Input and Output Data

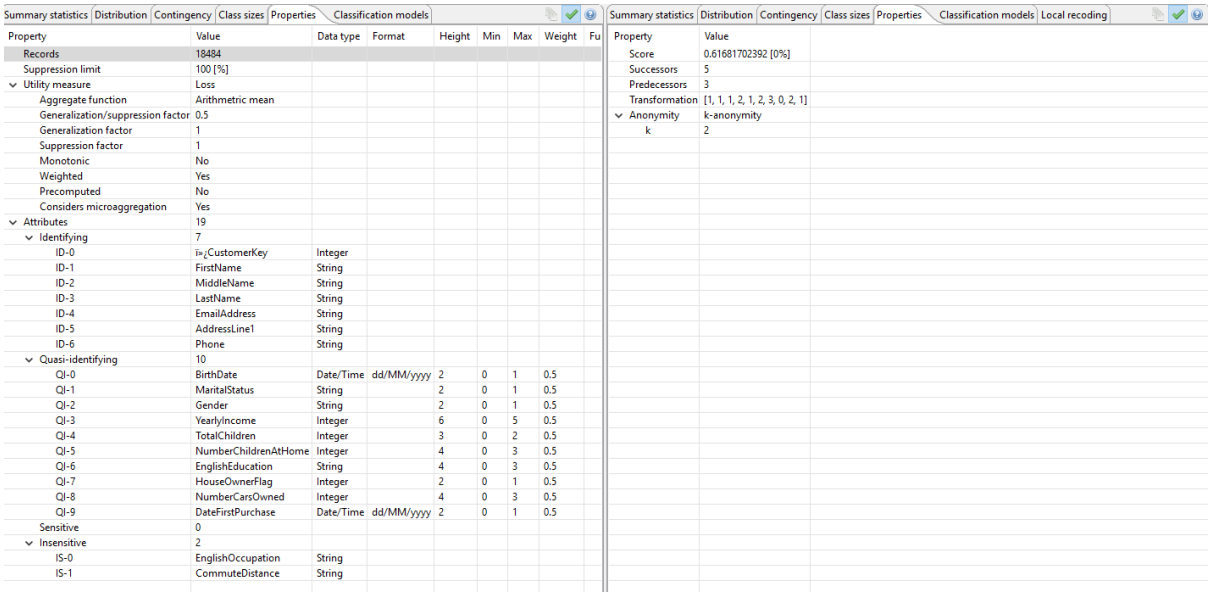


Figure 39: Properties of Input and Output Data

Properties of input data tab within ARX shows the basic properties of the input data set and the configuration used for anonymisation. The properties of output data tab show the basic properties about the selected data transformation as well as the resultant output data set. Figure 39 refers.

10.1.7 Classification Performance

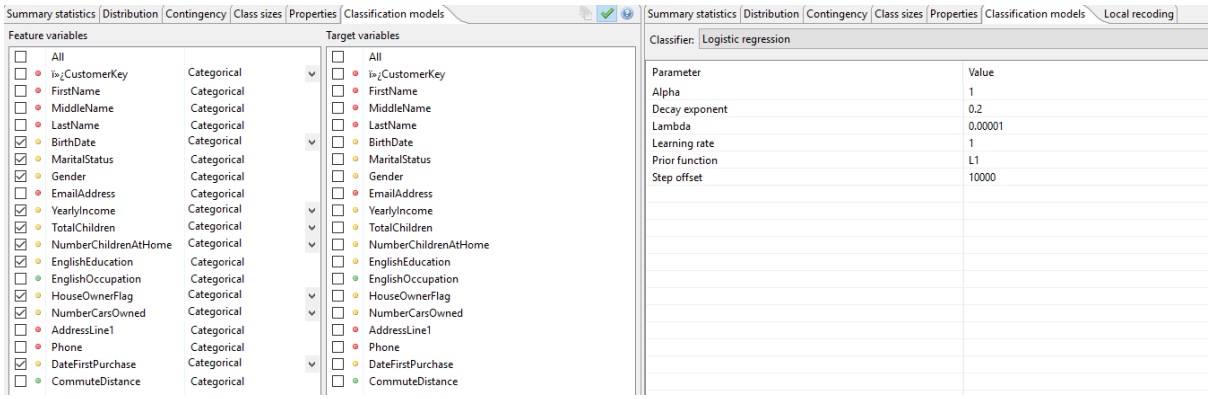
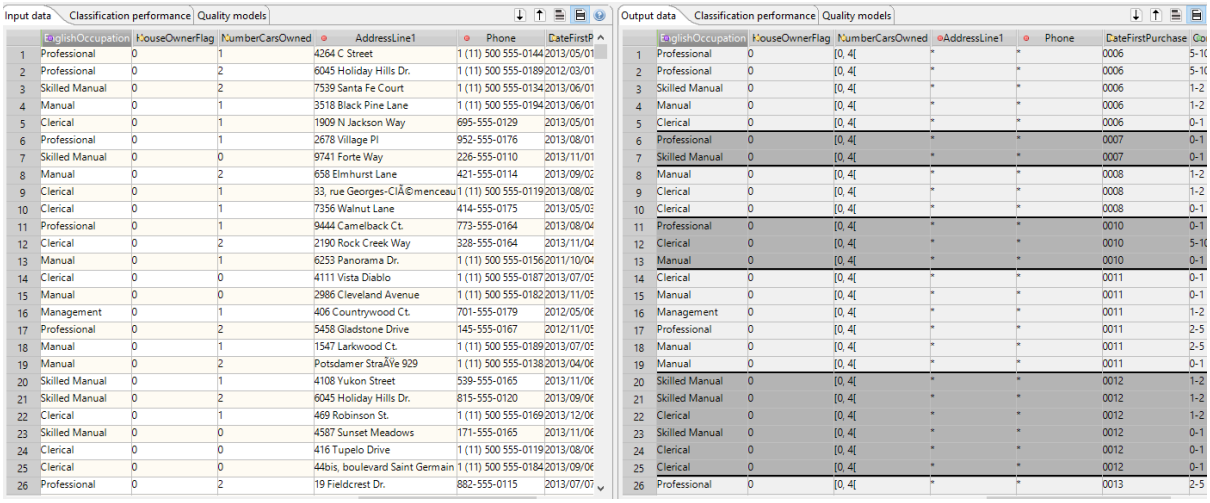


Figure 40: Classification Performance

Configuration of the classification models and their parameters can be done in this view. Performance of the models can also be seen. Please refer to the figure above.

10.2 L-diversity Data Utility Results

10.2.1 Input Output Data



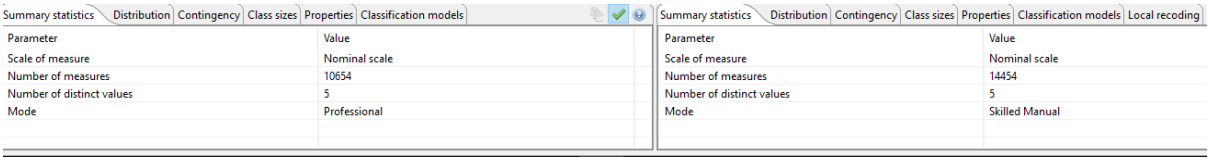
	EnglishOccupation	HouseOwnerFlag	NumberCarsOwned	AddressLine1	Phone	DateFirstP
1	Professional	0	1	4264 C Street	1 (11) 500 555-0144	2013/05/01
2	Professional	0	2	6045 Holiday Hills Dr.	1 (11) 500 555-0189	2012/03/01
3	Skilled Manual	0	2	7539 Santa Fe Court	1 (11) 500 555-0134	2013/06/01
4	Manual	0	1	3518 Black Pine Lane	1 (11) 500 555-0194	2013/06/01
5	Clerical	0	1	1909 N Jackson Way	695-555-0129	2013/05/01
6	Professional	0	1	2678 Village Pl	952-555-0176	2013/08/01
7	Skilled Manual	0	0	9741 Forte Way	226-555-0110	2013/11/01
8	Manual	0	2	658 Elmhurst Lane	421-555-0114	2013/09/02
9	Clerical	0	1	33, rue Georges-ClÃ©menceau	1 (11) 500 555-0119	2013/08/02
10	Clerical	0	1	7356 Walnut Lane	414-555-0175	2013/05/03
11	Professional	0	1	9444 Camelback Ct.	773-555-0164	2013/08/04
12	Clerical	0	2	2190 Rock Creek Way	328-555-0164	2013/11/04
13	Manual	0	1	6253 Panorama Dr.	1 (11) 500 555-0156	2011/10/04
14	Clerical	0	0	4111 Vista Diablo	1 (11) 500 555-0187	2013/07/05
15	Manual	0	0	2986 Cleveland Avenue	1 (11) 500 555-0182	2013/11/05
16	Management	0	1	406 Countrywood Ct.	701-555-0179	2012/05/06
17	Professional	0	2	5458 Gladstone Drive	145-555-0167	2012/11/05
18	Manual	0	1	1547 Larkwood Ct.	1 (11) 500 555-0189	2013/07/05
19	Manual	0	2	Potsdamer StraÃŸe 929	1 (11) 500 555-0138	2013/04/06
20	Skilled Manual	0	1	4108 Yukon Street	539-555-0165	2013/11/06
21	Skilled Manual	0	2	6045 Holiday Hills Dr.	815-555-0120	2013/09/06
22	Clerical	0	1	469 Robinson St.	1 (11) 500 555-0169	2013/12/06
23	Skilled Manual	0	0	4587 Sunset Meadows	171-555-0165	2013/11/06
24	Clerical	0	0	416 Tupelo Drive	1 (11) 500 555-0119	2013/08/06
25	Clerical	0	0	44bis, boulevard Saint Germain	1 (11) 500 555-0184	2013/09/06
26	Professional	0	2	19 Fieldcrest Dr.	882-555-0115	2013/07/07

	EnglishOccupation	HouseOwnerFlag	NumberCarsOwned	AddressLine1	Phone	DateFirstPurchase	Gr
1	Professional	0	[0, 4[	*	*	0006	5-1
2	Professional	0	[0, 4[	*	*	0006	5-1
3	Skilled Manual	0	[0, 4[	*	*	0006	1-2
4	Manual	0	[0, 4[	*	*	0006	1-2
5	Clerical	0	[0, 4[	*	*	0006	0-1
6	Professional	0	[0, 4[	*	*	0007	0-1
7	Skilled Manual	0	[0, 4[	*	*	0007	0-1
8	Manual	0	[0, 4[	*	*	0008	1-2
9	Clerical	0	[0, 4[	*	*	0008	1-2
10	Clerical	0	[0, 4[	*	*	0008	0-1
11	Professional	0	[0, 4[	*	*	0010	0-1
12	Clerical	0	[0, 4[	*	*	0010	5-1
13	Manual	0	[0, 4[	*	*	0010	0-1
14	Clerical	0	[0, 4[	*	*	0011	0-1
15	Manual	0	[0, 4[	*	*	0011	0-1
16	Management	0	[0, 4[	*	*	0011	1-2
17	Professional	0	[0, 4[	*	*	0011	2-5
18	Manual	0	[0, 4[	*	*	0011	2-5
19	Manual	0	[0, 4[	*	*	0011	0-1
20	Skilled Manual	0	[0, 4[	*	*	0012	1-2
21	Skilled Manual	0	[0, 4[	*	*	0012	1-2
22	Clerical	0	[0, 4[	*	*	0012	1-2
23	Skilled Manual	0	[0, 4[	*	*	0012	0-1
24	Clerical	0	[0, 4[	*	*	0012	0-1
25	Clerical	0	[0, 4[	*	*	0012	0-1
26	Professional	0	[0, 4[	*	*	0013	2-5

Figure 41: Input and Output Data

The input data set on shown on the left pane and the results after anonymisation is shown on the right pane.

10.2.2 Summary Statistics



Parameter	Value
Scale of measure	Nominal scale
Number of measures	10654
Number of distinct values	5
Mode	Professional

Parameter	Value
Scale of measure	Nominal scale
Number of measures	14454
Number of distinct values	5
Mode	Skilled Manual

Figure 42: Summary Statistics

The summary statistics for the chosen attribute is shown here.

10.2.3 Frequency Distribution

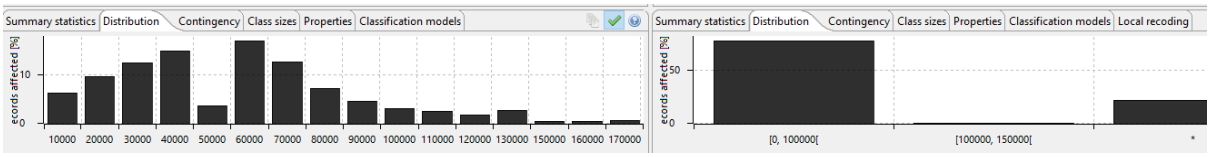


Figure 43: Frequency Distribution

A table or histogram that visualises the frequency distribution of the values of the selected attribute is shown above.

10.2.4 Contingency

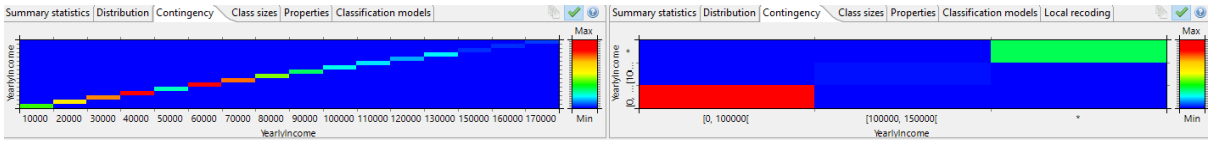


Figure 44: Contingency

The figure above represents a heat map that visually shows the contingency of two selected attributes. Contingency refers to the multivariate frequency distribution of the variables.

10.2.5 Equivalence Classes and Records

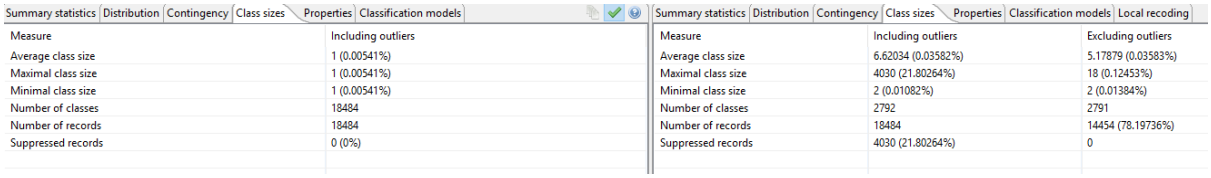


Figure 45: Equivalence Classes and Records

The figure above shows the equivalence classes and records which summarises the information and shows the minimal class size, maximal class size and average size of the equivalence classes.

10.2.6 Properties of Input and Output Data

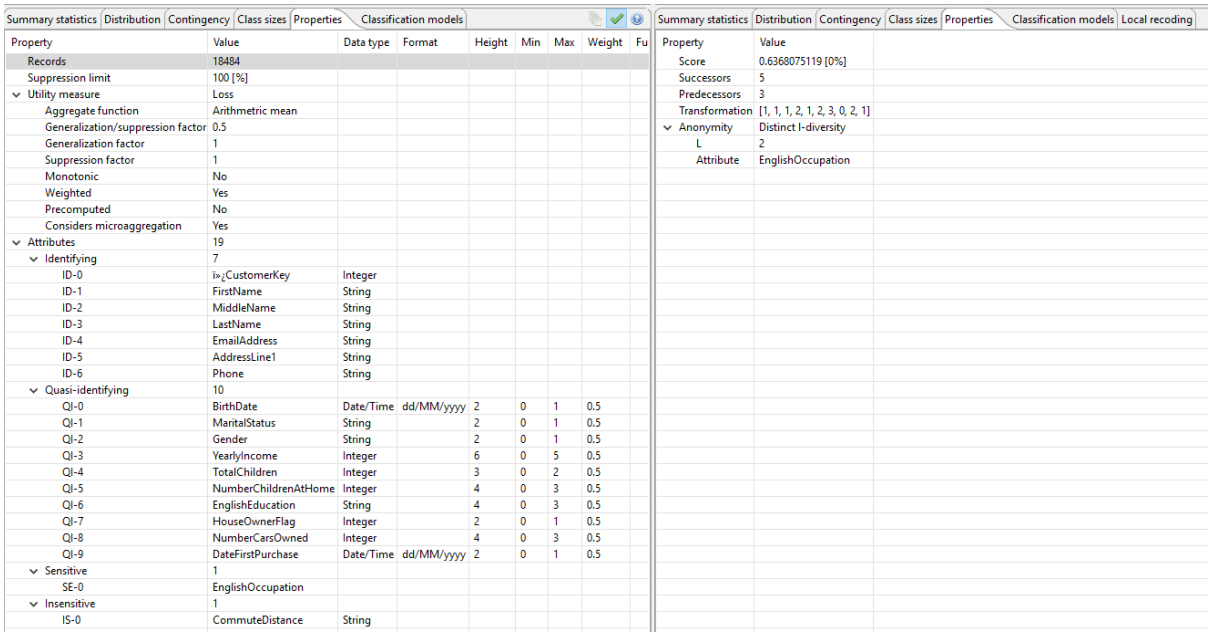


Figure 46: Properties of Input and Output Data

Properties of input data shows the basic properties of the input data set and the configuration used for anonymisation. The properties of output data tab show the basic properties about the selected data transformation as well as the resultant output data set.

10.2.7 Classification Performance

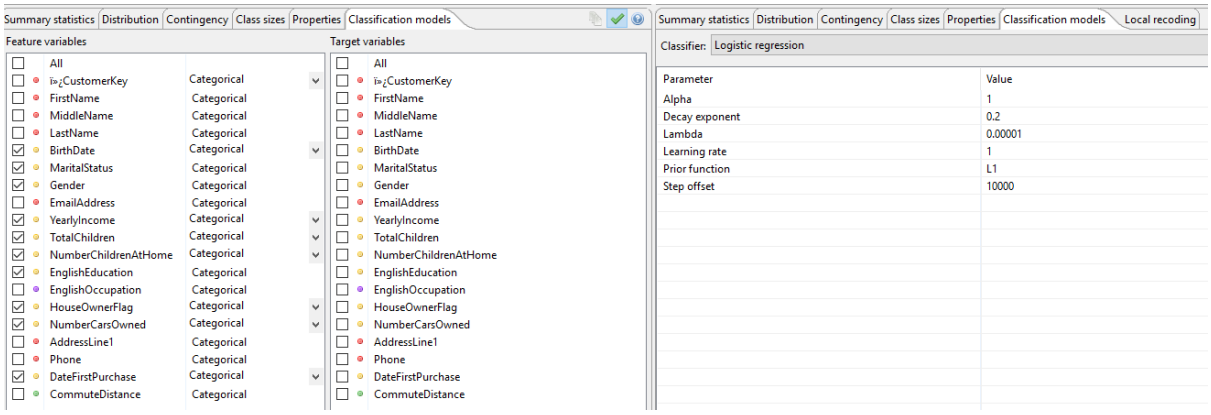


Figure 47: Classification Performance

Configuration of the classification models and their parameters can be done in this view. Performance of the models can also be seen. Please refer to the figure above.

10.3 K-anonymity Privacy Risks Results

10.3.1 Distribution of Risks

The distribution of risks between the input and output data set is shown side-by-side.

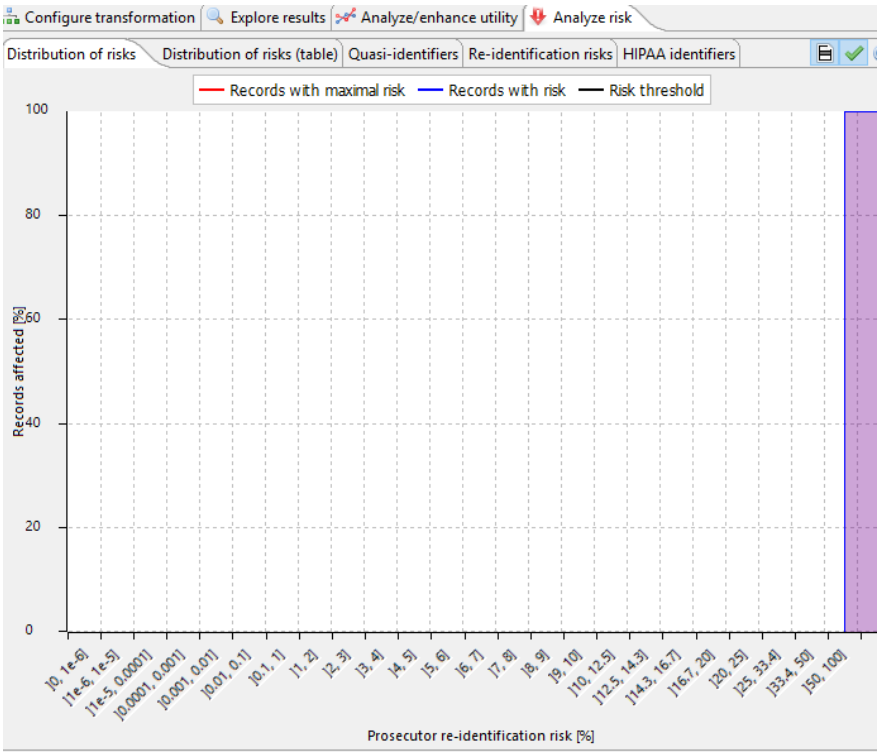
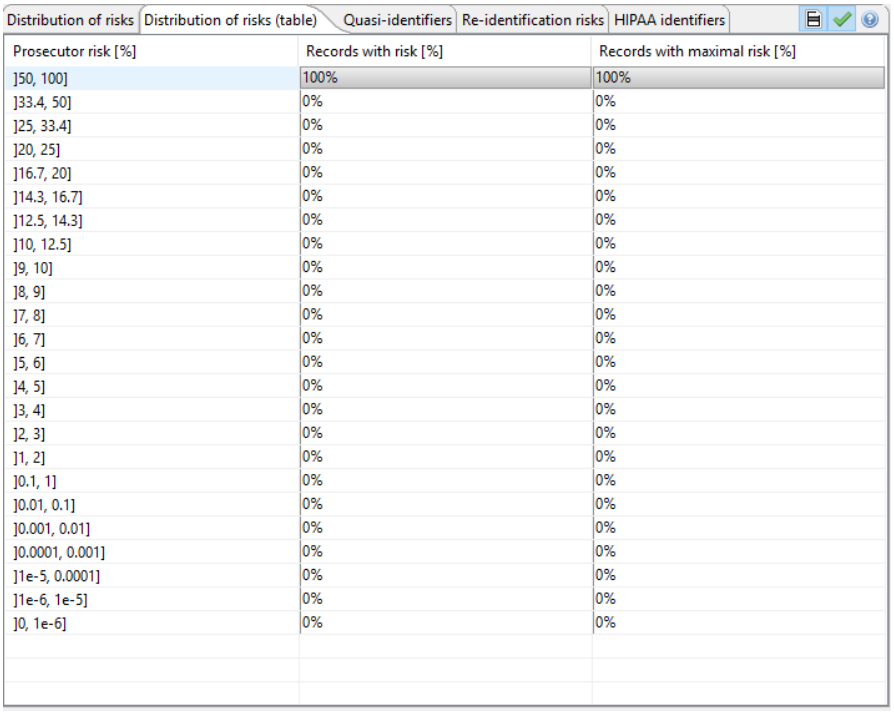


Figure 48: Distribution of Risks Input



Distribution of risks | Distribution of risks (table) | Quasi-identifiers | Re-identification risks | HIPAA identifiers

Prosecutor risk [%]	Records with risk [%]	Records with maximal risk [%]
]50, 100]	100%	100%
]33.4, 50]	0%	0%
]25, 33.4]	0%	0%
]20, 25]	0%	0%
]16.7, 20]	0%	0%
]14.3, 16.7]	0%	0%
]12.5, 14.3]	0%	0%
]10, 12.5]	0%	0%
]9, 10]	0%	0%
]8, 9]	0%	0%
]7, 8]	0%	0%
]6, 7]	0%	0%
]5, 6]	0%	0%
]4, 5]	0%	0%
]3, 4]	0%	0%
]2, 3]	0%	0%
]1, 2]	0%	0%
]0.1, 1]	0%	0%
]0.01, 0.1]	0%	0%
]0.001, 0.01]	0%	0%
]0.0001, 0.001]	0%	0%
]1e-5, 0.0001]	0%	0%
]1e-6, 1e-5]	0%	0%
]0, 1e-6]	0%	0%

Figure 49: Distribution of Risks in tabular form

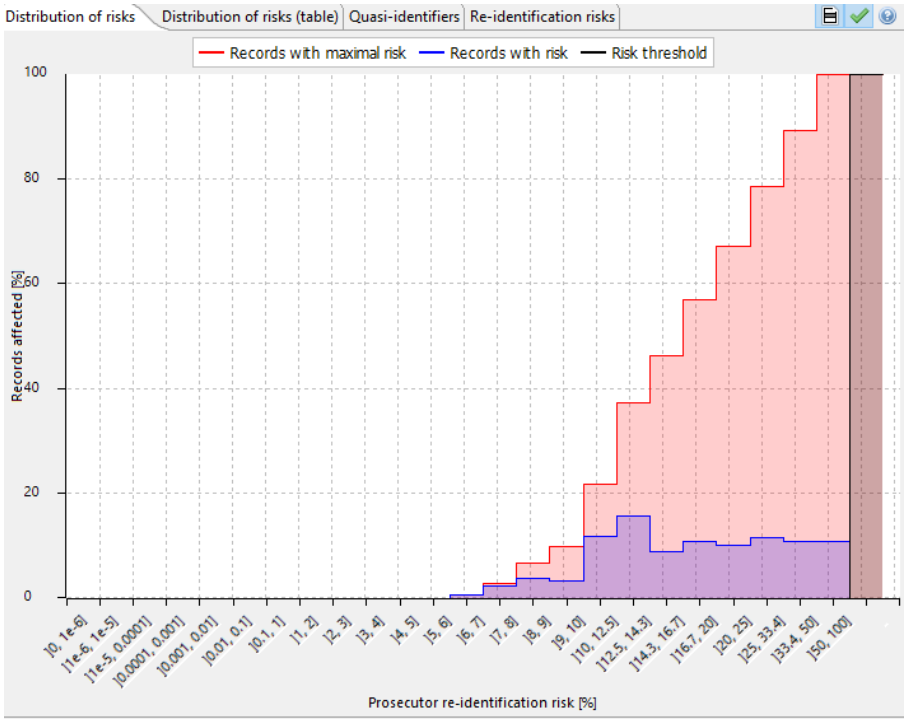
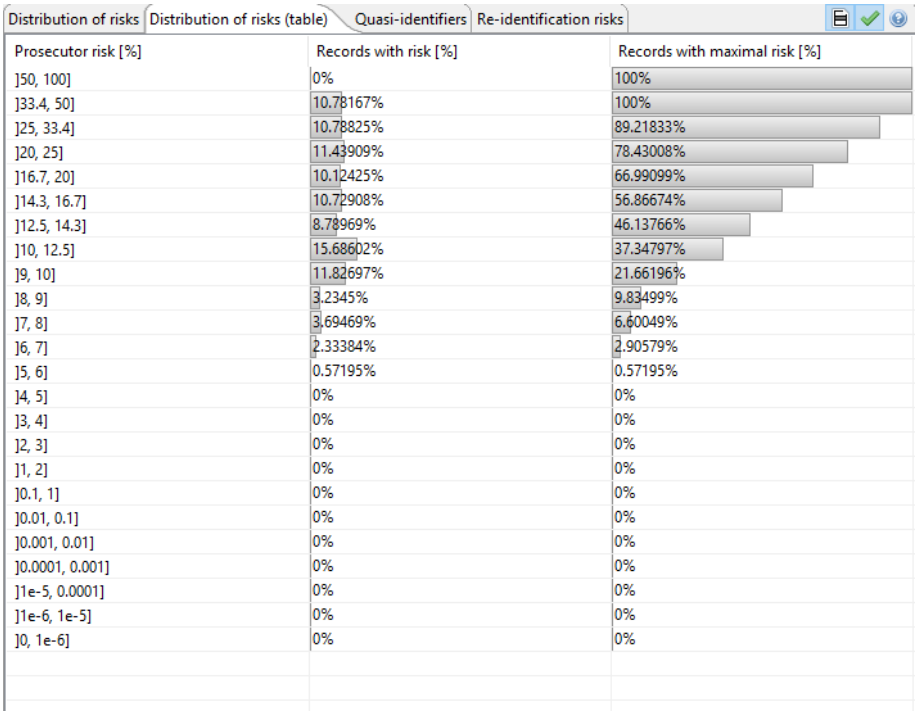


Figure 50: Distribution of Risks Output



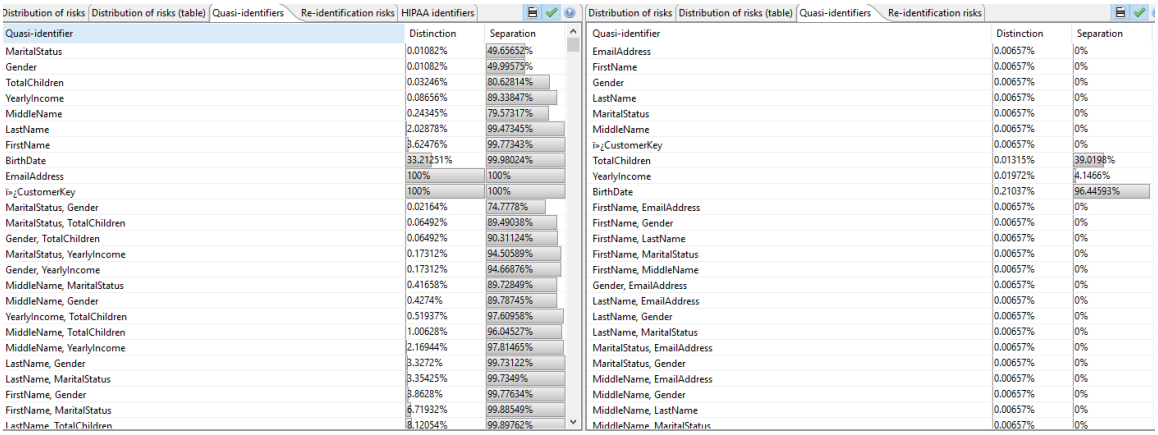
Distribution of risks | Distribution of risks (table) | Quasi-identifiers | Re-identification risks

Prosecutor risk [%]	Records with risk [%]	Records with maximal risk [%]
]50, 100]	0%	100%
]33.4, 50]	10.78167%	100%
]25, 33.4]	10.78825%	89.21833%
]20, 25]	11.43909%	78.43008%
]16.7, 20]	10.12425%	66.99099%
]14.3, 16.7]	10.72908%	56.86674%
]12.5, 14.3]	8.78969%	46.13766%
]10, 12.5]	15.68602%	37.34797%
]9, 10]	11.82697%	21.66196%
]8, 9]	3.2345%	9.83499%
]7, 8]	3.69469%	6.60049%
]6, 7]	2.33384%	2.90579%
]5, 6]	0.57195%	0.57195%
]4, 5]	0%	0%
]3, 4]	0%	0%
]2, 3]	0%	0%
]1, 2]	0%	0%
]0.1, 1]	0%	0%
]0.01, 0.1]	0%	0%
]0.001, 0.01]	0%	0%
]0.0001, 0.001]	0%	0%
]1e-5, 0.0001]	0%	0%
]1e-6, 1e-5]	0%	0%
]0, 1e-6]	0%	0%

Figure 51: Distribution of Risks in tabular form

10.3.2 Quasi-Identifiers

Within the quasi-identifiers tab, an analysis can be done with the combinations of attributes side-by-side. The associated re-identification risks can also be viewed. Please refer to Figure 52.



Distribution of risks | Distribution of risks (table) | Quasi-identifiers | Re-identification risks | HIPAA identifiers

Quasi-identifier	Distinction	Separation
MaritalStatus	0.01082%	49.65652%
Gender	0.01082%	49.99575%
TotalChildren	0.03246%	80.62814%
YearlyIncome	0.08656%	89.33847%
MiddleName	0.24345%	79.57317%
LastName	2.02878%	99.47345%
FirstName	3.62476%	99.77343%
BirthDate	33.21251%	99.98024%
EmailAddress	100%	100%
ï»¿CustomerKey	100%	100%
MaritalStatus, Gender	0.02164%	74.7778%
MaritalStatus, TotalChildren	0.06492%	89.49038%
Gender, TotalChildren	0.06492%	90.31124%
MaritalStatus, YearlyIncome	0.17312%	94.50589%
Gender, YearlyIncome	0.17312%	94.66876%
MiddleName, MaritalStatus	0.41658%	89.72849%
MiddleName, Gender	0.4274%	89.78745%
YearlyIncome, TotalChildren	0.51937%	97.60958%
MiddleName, TotalChildren	1.00628%	96.04527%
MiddleName, YearlyIncome	2.16944%	97.81465%
LastName, Gender	3.3272%	99.73122%
LastName, MaritalStatus	3.35425%	99.7349%
FirstName, Gender	3.8628%	99.77634%
FirstName, MaritalStatus	6.71932%	99.88549%
LastName, TotalChildren	8.12054%	99.89762%

Distribution of risks | Distribution of risks (table) | Quasi-identifiers | Re-identification risks

Quasi-identifier	Distinction	Separation
EmailAddress	0.00657%	0%
FirstName	0.00657%	0%
Gender	0.00657%	0%
LastName	0.00657%	0%
MaritalStatus	0.00657%	0%
MiddleName	0.00657%	0%
ï»¿CustomerKey	0.00657%	0%
TotalChildren	0.01315%	39.0198%
YearlyIncome	0.01972%	4.1466%
BirthDate	0.21037%	96.44593%
FirstName, EmailAddress	0.00657%	0%
FirstName, Gender	0.00657%	0%
FirstName, LastName	0.00657%	0%
FirstName, MaritalStatus	0.00657%	0%
FirstName, MiddleName	0.00657%	0%
Gender, EmailAddress	0.00657%	0%
LastName, EmailAddress	0.00657%	0%
LastName, Gender	0.00657%	0%
LastName, MaritalStatus	0.00657%	0%
MaritalStatus, EmailAddress	0.00657%	0%
MaritalStatus, Gender	0.00657%	0%
MiddleName, EmailAddress	0.00657%	0%
MiddleName, Gender	0.00657%	0%
MiddleName, LastName	0.00657%	0%
MiddleName, MaritalStatus	0.00657%	0%

Figure 52: Quasi-identifiers

10.4 L-diversity Privacy Risks Results

10.4.1 Distribution of Risks

The distribution of risks between the input and output data set is shown side-by-side.

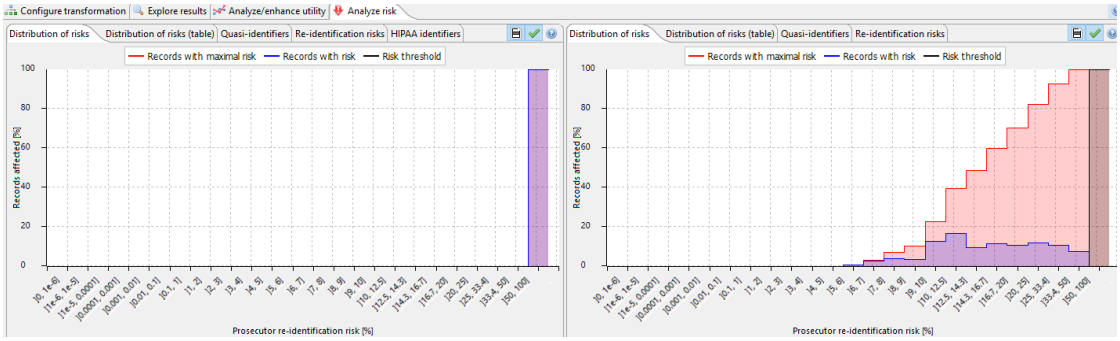


Figure 53: Distribution of Risks

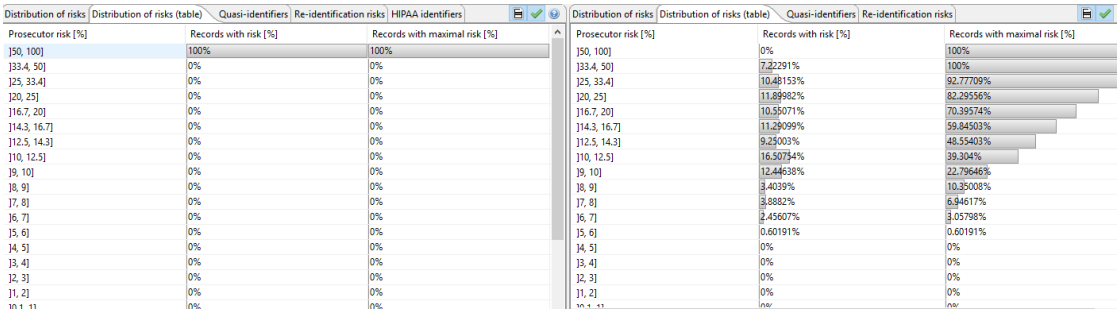


Figure 54: Distribution of Risks in tabular form

10.4.2 Quasi-identifiers

Within the quasi-identifiers tab, an analysis can be done with the combinations of attributes side-by-side. The associate re-identification risks can also be viewed.

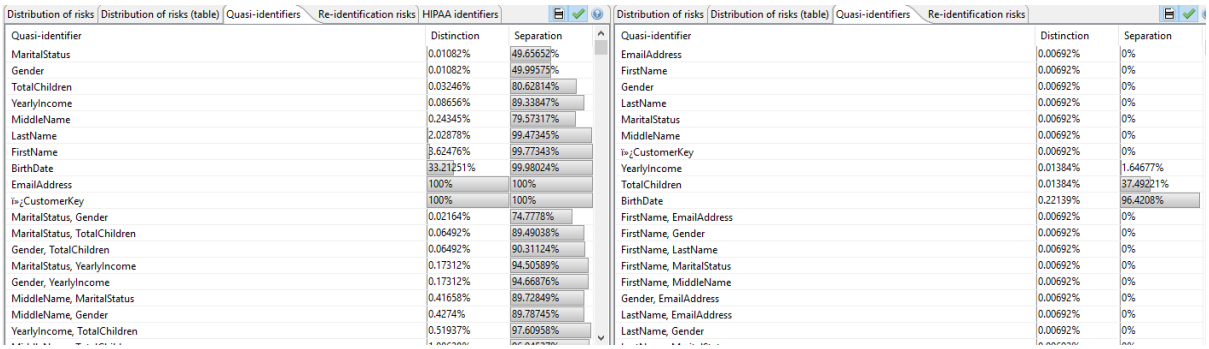
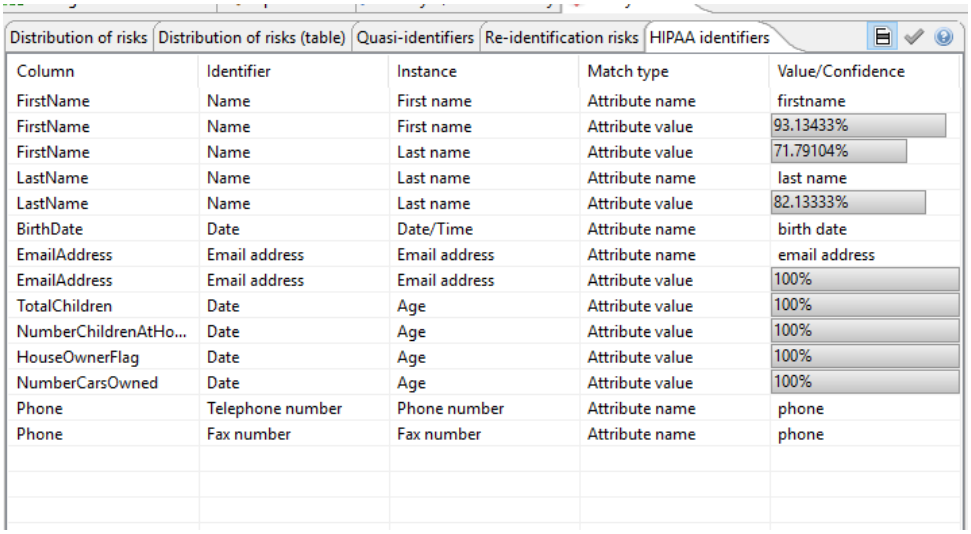


Figure 55: Quasi-identifiers

10.5 HIPAA Tab



Column	Identifier	Instance	Match type	Value/Confidence
FirstName	Name	First name	Attribute name	firstname
FirstName	Name	First name	Attribute value	93.13433%
FirstName	Name	Last name	Attribute value	71.79104%
LastName	Name	Last name	Attribute name	last name
LastName	Name	Last name	Attribute value	82.13333%
BirthDate	Date	Date/Time	Attribute name	birth date
EmailAddress	Email address	Email address	Attribute name	email address
EmailAddress	Email address	Email address	Attribute value	100%
TotalChildren	Date	Age	Attribute value	100%
NumberChildrenAtHo...	Date	Age	Attribute value	100%
HouseOwnerFlag	Date	Age	Attribute value	100%
NumberCarsOwned	Date	Age	Attribute value	100%
Phone	Telephone number	Phone number	Attribute name	phone
Phone	Fax number	Fax number	Attribute name	phone

Figure 56: HIPAA Tab and Identifiers

The HIPAA tab within ARX automatically detects the eight identifiers that must be modified or removed from a data set.