Erasmus University Rotterdam

School of Economics

Master Thesis – Data Science & Marketing Analytics

Privacy Preserving Data Publishing Techniques: a Comparative Analysis with Medical Records

---

Author: Frédérique P. J. Smulders
Student ID: 505724
Erasmus Supervisor: prof. dr. Martijn G. de Jong
Second Assessor: dr. Carlo C. Cavicchia
Date: July 5th, 2024

**The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics, or Erasmus University Rotterdam*

Abstract

This thesis explores privacy-preserving data publishing (PPDP) techniques, focusing on k-anonymization and differential privacy, and their effectiveness in managing privacy, data utility, and computational complexity within tabular medical records. As data breaches become more frequent in our data-driven society, robust methods to protect sensitive information are crucial, especially in light of strict regulations like the General Data Protection Regulation (GDPR).

Through the use of generalization techniques, k-anonymization guarantees that each individual cannot be distinguished from at least k-1 other individuals, while differential privacy ensures that the inclusion or exclusion of any individual's data does not substantially affect the analysis results by adding noise. The study implements k-anonymization using the Multidimensional Mondrian algorithm and differential privacy using the Iterative Proportional Fitting (IPF) algorithm, revealing the expected trade-off between privacy protection and data utility, where lower privacy levels result in higher data utility and vice versa.

The sparsity in the original data posed challenges for k-anonymization as the strict partitioning technique was unable to generate a dataset that satisfied the k-anonymity criterion. Nonetheless, the relaxed partitioning technique functioned effectively, sometimes necessitating more aggressive generalization and at other times none at all to achieve higher levels of privacy. For initial small increases in the privacy budget of the IPF method, the utility gain exceeded the privacy loss, making the higher privacy budget an appealing choice for data publishers. Contrary to general assumptions in literature, the findings reveal that this specific differential privacy implementation is more computationally efficient than the k-anonymization implementation, making it better suited for handling large datasets.

Considering these findings and the inherent characteristics of both methods, the study recommends using the IPF algorithm with differential for sharing datasets, as it maintains a similar data structure with high data usability, flexibility, provides robust privacy guarantees, and is computationally efficient. For real-world applications, the results in this thesis also guide the selection of appropriate parameters and stimulates further exploration into efficient and responsible data sharing practices.

Contents

1 Introduction **3**

1.1 Scope . . . 3

1.2 Contribution . . . 5

2 Literature Review **7**

2.1 Terminology . . . 7

2.1.1 Table Components . . . 8

2.1.2 Privacy Goals and Attacks . . . 9

2.1.3 Anonymization Operations . . . 11

2.2 Privacy Models . . . 14

2.2.1 K-anonymization . . . 14

2.2.2 Differential Privacy . . . 15

3 Methodology **18**

3.1 Data Collection . . . 18

3.2 Inter-methodology metrics . . . 18

3.2.1 Verification statistical assumptions . . . 19

3.3 Multidimensional Mondrian K-anonymization . . . 21

3.3.1 Process . . . 21

3.3.2 Evaluation Metrics . . . 23

3.3.3 Challenges . . . 24

3.4 Iterative Proportional Fitting with Differential Privacy . . . 24

3.4.1 Process . . . 24

3.4.2 Evaluation Metrics . . . 27

3.4.3 Challenges . . . 28

4 Results **29**

4.1 Mondrian k-anonymization . . . 29

4.1.1 Strict partitioning . . . 29

4.1.2 Relaxed partitioning . . . 31

4.2 IPF with Differential Privacy . . . 32

4.2.1 Non-DP synthesis . . . 32

4.2.2 DP syntheses . . . 34

4.3 Logistic Regression Analysis . . . 37

4.3.1 Benchmark model . . . 37

4.3.2 Mondrian model . . . 39

4.3.3 IPF with DP model . . . 41

4.3.4 Comparison . . . 42

5 Discussion **44**
5.1 Expected findings 44
5.2 Unexpected findings 45
5.3 Limitations 46

6 Conclusion **48**

7 References **51**

8 Appendix **54**

1 Introduction

In recent years, the development of data-driven technologies and the widespread adoption of machine learning algorithms have revolutionized various aspects of society, promising remarkable advantages in areas ranging from healthcare to the military. The ability to process vast amounts of data and extract meaningful insights has led to groundbreaking advancements, driving innovation and progress across industries. However, this rapid expansion of data-driven technologies also came with significant challenges, particularly in terms of privacy.

One of the key issues resulted from the rise of data-driven technologies is the increasing amount of privacy leaks and composition attacks. These threats to privacy occur when seemingly harmless pieces of data are combined to reveal sensitive information about individuals. A famous example is the case of Netflix, where they published a dataset containing information about thousands of subscribers for the purpose of a Netflix Prize contest (an open competition to find the best algorithm to predict user ratings for movies). However, contestants quickly found out that the privacy of subscribers included in the revealed data set was exposed as the seemingly anonymous movie ratings could be cross-referenced with public data sets to identify individual users. Narayanan and Shmatikov (2008) concluded that knowing only 5 to 10 background attributes was enough to de-anonymize individuals in a dataset. Such incidents underscore the need for robust privacy protection mechanisms, especially for situations in which the information affects the individual's safety.

In response to these ethical threats, regulators have authorized legislation aimed at safeguarding individual privacy rights. In 2018, the General Data Protection Regulation (GDPR) was established to strengthen data protection and privacy for individuals within the European Union (EU) and the European Economic Area (EEA). The GDPR sets strict requirements on organizations regarding the collection, processing, and storage of personal data with substantial fines in case of non-compliance. The new law emphasizes the need for anonymization methods which make it impossible to identify individuals considering factors such as time, costs, and current and future technological developments within rationality. While the factor of future technological development limits the possibilities for organizations greatly, the rationality factor will enable them to leverage data by adopting intelligent solutions.

1.1 Scope

Europe is a pioneer in enacting strengthened privacy laws, but it is not alone in this effort. Countries worldwide are rapidly responding to growing privacy challenges by imitating Europe's regulatory framework. To offer a comprehensive overview of this global landscape, this research focuses on an analysis of methods utilizing data from diverse countries, specifically France, the United States, Brazil, and Singapore.

As highlighted by Kaissis et al. (2020), there are distinct realms of techniques aimed at either

protecting data or safeguarding algorithms. For instance, techniques designed to protect algorithms include methods to prevent model inversion attacks. In a model inversion attack, attackers can potentially reconstruct sensitive data from the model's parameters by strategically providing various synthetic or random inputs to the model and observing the outputs to infer the underlying data patterns. Preventative techniques such as adversarial training include adversarial examples designed to mimic potential inversion attacks in the training process and thereby train the model to resist such reverse-engineering attempts (Prakash et al., 2020). However, this study emphasizes evaluating methods that allow for the safe and compliant release of complete tabular datasets, rather than protecting the integrity and confidentiality of the algorithms, and to protect it from unauthorized access or disclosure.

Moreover, as this thesis focuses exclusively on releasing datasets where the identity of individuals is protected and the usability of the data is still good, it falls into the category of Privacy Preserving Data Publishing (PPDP) methods. The core objective is to explore how sensitive information can be anonymized and securely shared with external parties. In contrast, Privacy Preserving Data Mining (PPDM) involves extracting useful information and patterns from data during internal data analysis and processing (Fung et al., 2010). A well-known example of such a technique is differentially private stochastic gradient descent (DP-SGD) which adds noise to the gradients calculated during each step of model training, thereby protecting the privacy of individual data points in the training dataset (Song et al., 2013). While both PPDP and PPDM are crucial for safeguarding privacy, this study is confined to the strategies and implementations relevant to the secure publication of data.

As briefly mentioned, PPDP methods anonymize tabular datasets to securely share them with external parties. This involves modifying the data to reduce re-identification risks prior to sharing. This approach is distinctly different from encryption techniques, which focus on maintaining data confidentiality by encrypting the data with a cipher key, allowing only entities with that key to access the original data (Fung et al., 2010). For example, Homomorphic Encryption (HE) enables third parties to perform computations on encrypted data without needing to decrypt it, thus maintaining confidentiality throughout the data analysis process (Gentry, 2009). Since encryption techniques do not require actual modifications to the data, they fall outside the scope of this thesis, which is focused on methods for securely publishing anonymized datasets.

Various PPDP methods have been proposed by researchers with Fung et al. (2010) providing an extensive overview within this scope. Among these methodologies, k-anonymization is one of the most foundational anonymization algorithms (Majeed and Lee, 2020). This algorithm ensures that if an individual possesses a specific value as a quasi-identifier, at least $k-1$ other observations must share the same value (Samarati and Sweeney, 1998). By enforcing this condition, k-anonymization mitigates the risk of re-identification through linking attacks, thereby limiting adversaries' ability to uncover sensitive information about individuals.

K-anonymization techniques can be applied to multidimensional datasets with relatively modest computational overhead, rendering them viable for real-world scenarios. However, datasets in those

scenarios sometimes also exhibit high dimensionality and complex structures. The combination is argued to pose challenges for traditional k-anonymization approaches. High dimensionality complicates the selection of an appropriate k-value as it often results in either too much loss of information or too high privacy risk. Consequently, the need arises to compare k-anonymization with other, more recently developed methodologies to understand their respective strengths and weaknesses.

As first presented by Dwork (2006), the Differential Privacy (DP) framework protects individual privacy by guaranteeing the outcome of an analysis remains unaffected by the presence or absence of any single observation. Since the debut of this theoretical framework, many researchers have utilized it by combining it with various forms of machine learning techniques. For example, Xiao et al. (2010) underscored the efficacy of differential privacy in preserving privacy through the addition of noise drawn from the Laplace distribution to aggregated range-count values, thereby obfuscating individual records before publishing the resulting anonymized dataset.

The framework provides a strong and quantifiable guarantee, offering plausible deniability for the inclusion of individuals in the dataset. One of the key advantages of this technique is the resilience against various attacks, including linking attacks and membership inference attacks. However, the introduction of noise or data perturbation may potentially compromise data utility, demanding a careful balance between privacy and utility considerations. Moreover, the differential privacy framework is complex, and determining the appropriate amount of noise to abide by the framework requires precise calculations based on the sensitivity of the data and may incur computational overhead.

1.2 Contribution

As underscored by Fung et al. (2010), the expanding gap between privacy-threatening technologies and the adoption of privacy-preserving measures requires attention. Thus, this study attempts to enrich academic literature by conducting an assessment of the two PPDP techniques, delving into their respective strengths, limitations, and trade-offs. Such an analysis promises to inform more reasonable decision-making with regard to a suitable method and parameter settings in real-world contexts.

Moreover, the comprehensive examination of both intra-methodology and inter-methodology metrics provides valuable insights that aid in selecting the most suitable privacy-preserving method. Intra-methodology metrics, specific to each anonymization technique, result in a decision-making guideline. This guidance helps a data publisher select a technique based on their preferences regarding dataset characteristics, the utility-privacy trade-off, and computational complexity. More importantly, the intra-methodology metrics assist a data publisher in selecting the optimal parameter settings.

Inter-methodology metrics enable a data publisher to compare the practicality of the two techniques and in turn contribute to the choice of method. Specifically, the impact on data utility will be

analyzed by comparing the prediction accuracies of the original data to those of the anonymized datasets. Additionally, the execution times of the anonymization steps for each approach will be compared to assess computational efficiency.

Beyond the academic contributions, this comparative analysis of PPDP methodologies is significant for fostering innovation across various disciplines. The adoption of efficient privacy techniques enables more ethical and responsible data sharing, which in turn facilitates data-driven research, stimulates innovation, and encourages interdisciplinary collaboration. This is particularly beneficial in the healthcare domain, where the GDPR legislation significantly restricts the potential for new medical AI techniques due to the high sensitivity in medical records (Kaissis et al., 2020). By adopting efficient privacy measures, it becomes possible to drive the development of novel treatments, therapies, and healthcare solutions.

In order to bridge the gap highlighted by both the contribution to academics and innovation, I will implement k-anonymization and differential privacy methods on a dataset containing sensitive information about individuals concerning the illegal purchasing of prescription drugs without prescription and thereby addressing the following research questions:

How do k-anonymization and differential privacy methods compare in terms of preserving privacy, maintaining data utility, and managing computational complexity when applied to sensitive healthcare datasets?

The remainder of this thesis will provide an answer to this research question by first outlining a comprehensive literature review in chapter 2. This chapter will examine the theoretical background of the subject as well as existing research on PPDP methods with a focus on k-anonymization and differential privacy. Chapter 3 will present the selected methods employed in this study, including the details of the data set, the mathematics behind the chosen method, and the evaluation metrics used to assess the effectiveness of these methods. The results will be presented and discussed in chapter 4, highlighting the strengths, limitations, and trade-offs of each approach. Chapter 5 will offer a discussion on the expected findings, the unexpected findings, the limitations of the research while also giving recommendations for future research. And finally, chapter 6 will provide conclusions drawn from the findings of the study and provide recommendations for potential stakeholders.

2 Literature Review

In this section, I aim to explain the contemporary academic landscape of PPDP methods. To understand the topic, I will first describe the key terminology and concepts in this research domain. Subsequently, I will present relevant theories and models that underpin this research by clarifying empirical evidence and findings from previous studies.

2.1 Terminology

There are distinct phases involved in handling data, which can be simplified into three main stages: data collection, data publishing, and data mining as outlined by Fung et al. (2010) and the flowchart in Figure 1. During the data collection phase, a data publisher gathers information from record owners. In the realm of medical records for insurance purposes, for instance, the data publisher may be a hospital or clinic, while the record owner is typically the insured individual. Then in the data publishing phase, the collected data from the insured party (such as medical history, treatments, and health habits) is stored and subsequently shared with a recipient for the data mining stage. In the scenario described, the data recipient could be a medical insurer utilizing the data to construct models or analyze trends. Moreover, a data publisher might also opt to release the data to the public for research, educational, or awareness-raising purposes.

Figure 1: Data Handling Phases

Note. A framework of the three stages in data handling with its respective main stakeholders.

In the data publishing phase, two distinct scenarios emerge. Firstly, when the data publisher is deemed untrustworthy, there exists a risk of potential attempts to extract sensitive information from record owners. In such cases, ensuring anonymity during the data collection phase becomes imperative to preempt any privacy breaches by the data recipient. Conversely, in scenarios where

the data publisher is considered trustworthy, record owners may willingly provide personal and sensitive information (Fung et al., 2010). Within the scope of this thesis, only scenarios where the data publisher is trustworthy are examined, given the focus on PPDP techniques.

Even with a trustworthy data publisher, this trust is not necessarily extended to the data recipient. Even if the recipient is a reputable entity, such as a well-established medical insurance company, the potential for malicious intent among one of the employees with access to the data cannot be disregarded. Additionally, it is often unknown who the recipient will be at the time of data sharing, as the data might be shared with the public. Hence, prior to data sharing, the data publisher must ensure robust privacy protection for individuals included in the dataset.

2.1.1 Table Components

In this section, I'll elaborate on the components of the table, drawing upon insights from Mejeed and Lee (2020), as well as Fung et al. (2010). Each record within tabular data, commonly referred to as a tuple, comprises several distinct components, as detailed below.

As illustrated in the formula below, a private table D consists of attributes categorized into direct identifiers (DIs), quasi-identifiers (QIs), sensitive attributes (SAs), and non-sensitive attributes (NSAs). It's important to note that these attributes are mutually exclusive.

$$D(\text{Direct Identifier}, \text{Quasi Identifier}, \text{Sensitive Attribute}, \text{Non Sensitive Attribute})$$

Direct identifiers (DIs) are attributes that directly facilitate the identification of specific individuals and are thus considered highly sensitive. Examples of DIs include names, social security numbers, email addresses, and phone numbers. Prior to the anonymization process, DIs are entirely removed from D as they do not contribute valuable information for analysis.

Quasi-identifiers (QIs), while less sensitive than DIs, still pose privacy risks. QIs alone do not identify individuals, but when combined with other QIs from auxiliary information, they can potentially reveal individual identities. Common QIs include date of birth, gender, ZIP code, race, and occupation. Research by Narayanan and Shmatikov (2008) demonstrated the ability to de-anonymize datasets with as few as 5 to 10 QIs from auxiliary data, emphasizing the privacy risks associated with QIs. Consequently, anonymization operations are applied to QIs during the anonymization process to strike a balance between preserving valuable information for data mining purposes and mitigating the risk of attacks—an issue commonly referred to as the anonymization problem.

Sensitive attributes (SAs) demand an even higher level of protection due to their potential impact on individuals' privacy and well-being when exposed. SAs encompass sensitive or private information related to financial or medical circumstances, such as medical diagnoses, sexual orientation, criminal records, or in our case acknowledgment of unauthorized prescription drug use. To preserve the

high informativeness and thus the utility of the data for analytical purposes, methods such as k-anonymization sometimes retain SAs in their original form and only modifying the QIs to ensure privacy for both the QIs and the SA.

Non-sensitive attributes (NSAs) contain no personally identifiable or sensitive information and are utilized for analysis or processing purposes. Examples of NSAs include product categories and timestamps. Given their low risk of de-identification, NSAs are published without modification.

Ultimately, the data publisher shares an anonymous table (T) in the format specified in the formula below, wherein QI* represents the anonymized version of QI after anonymization operations conducted on QI within D.

$$T(QI^*, SensitiveAttribute, NonSensitiveAttribute)$$

2.1.2 Privacy Goals and Attacks

Privacy by anonymization has long been a paramount concern in data analysis, dating back to the fundamental work of Dalenius in 1977. Dalenius concluded that true privacy in databases necessitates that no one should be able to learn anything about an individual without access to the database. However, as Dwork (2006) proved with the Fundamental Law of Information Recovery, achieving such an unconditional level of privacy is impossible due to all sorts of privacy threats promised on table T. Researchers Majeed and Lee (2020) summarized the privacy attacks in subcategorizations and called them identity disclosure, attribute disclosure, and membership disclosure.

Identity disclosure occurs when an attacker successfully identifies a specific individual within a supposedly anonymized dataset by cross-referencing the remaining attributes with external data sources. A simple example is shown by Tables 1 and 2. A hospital supposedly anonymized the records (Table 1) and published the resulting table (Table 2). If an attacker cross-references the anonymized dataset with the external information that a person living in ZIP code 12347 who is 28 years old and suffers from hypertension, he can uniquely identify Jim Brown. Hence, removing only direct identifiers is not always sufficient and can end up compromising someone's identity. Besides the simple example and the Netflix breach, a notable instance of an identity disclosure breach includes the infamous case of AOL's release of anonymized search queries. After the online service provider removed the direct identifiers, it was later cross-referenced with public data by reporters from the New York Times, leading to the identification of individuals like Thelma Arnold (Ohm, 2009).

Table 1: Original hospital records

Name	Age	ZIP_Code	Disease
John Smith	45	12345	Flu
Jane Doe	34	12346	Diabetes
Jim Brown	28	12347	Hypertension
Emily White	50	12348	Asthma

Note: This table presents original hospital records including the name, age, ZIP code, and disease of each individual.

Table 2: Anonymized hospital records

Age	ZIP_Code	Disease
45	12345	Flu
34	12346	Diabetes
28	12347	Hypertension
50	12348	Asthma

Note: This table presents anonymized hospital records without the DI (Name), but age, ZIP code, and disease information remains.

On the other hand, attribute disclosure occurs when sensitive information within the SA is linked to a specific individual, often exploiting imbalances in datasets as they lack heterogeneity. Again, a hospital publishes records without the direct identifiers. This time, the attacker knows that Emily White lives in ZIP code 12348 and is 50 years old. The combination of information means the attacker can deduce the sensitive attribute: her health condition. In 1997, the Massachusetts Group Insurance Commission did something similar and released hospital records that were subsequently cross-referenced with voter registration data, resulting in the reidentification of a governor's medical record stating his disease (Ohm, 2009).

Membership disclosure poses a different threat, wherein attackers can infer the presence of individuals in the dataset T without directly identifying them. There might be a situation in which an attacker wants to confirm if Jane Doe who he knows is 34 years old and lives in ZIP code 12346 has been treated at the hospital. Once the attacker observes the external information, he knows that the anonymized dataset (Table 2) includes this information and can thus confirm the person being a patient. While the identities or attributes remain undisclosed, this form of disclosure can still jeopardize individuals, as also exemplified by the case study conducted by Garner and Kim (2019) on DNA ancestry databases. They identified individuals' membership in the database from the companies 23andMe and AncestryDNA and exposed users' sensitive health data which left them vulnerable to discrimination or exploitation.

Recognizing the infeasibility of Dalenius's privacy goal due to potential privacy attacks, a more adaptable approach has emerged, aiming to determine general trends without compromising indi-

viduals' private information. Academic literature has explored various methodologies to address these evolving privacy concerns which will be discussed in depth in Section 2.2.

The evolution of privacy goals has resulted in a shift in definition of privacy as well. In this research, data privacy is defined as an individual's ability to exert control over their personal data (*What Is Data Privacy? | IBM*, n.d.). According to GDPR regulations (*Art. 4 GDPR – Definitions - General Data Protection Regulation (GDPR)*, 2018), personal data contains any information related to an identified or identifiable natural person. Identification can occur directly or indirectly through identifiers such as names, identification numbers, or characteristic factors of the individual.

2.1.3 Anonymization Operations

To fulfill the privacy requirements of the set privacy goal, various anonymization operations can be implemented collectively or independently on the original dataset D. These techniques include generalization, suppression, anatomization, permutation, and perturbation, each serving distinct purposes. While there exist various versions of these operations, I will provide a general overview of their concepts for clarity.

As the term already reveals, generalization involves replacing specific values in a dataset with broader or less precise values. Often applied to quasi-identifiers, this technique aims to prevent linking attacks with auxiliary information while preserving the dataset's overall structure and utility for data mining purposes. For instance, replacing specific ages (e.g. 25, 32, 35) with age ranges (e.g. 20-30, 30-40, 30-40) obscures individual ages while retaining information about age distributions. Thus, an anonymized by generalization table based on Table 2 could result in Table 3, where the columns referring to age and ZIP code are changed.

Table 3: Generalized Hospital Records

Age_Range	ZIP_Code_Range	Disease
40-49	12340-12349	Flu
30-39	12340-12349	Diabetes
20-29	12340-12349	Hypertension
50-59	12340-12349	Asthma

Note: This table shows generalized hospital records where age and ZIP code have been generalized to ranges.

Suppression conceals details within QIs as well, but by either removing or masking data elements entirely. This operation can involve suppressing entire rows (Le Fevre et al., 2005), instances of specific values (Wang et al., 2007), or certain occurrences of a specific value in a table (Cox, 1980). Examples include masking certain ages or ZIP codes as observed in Table 4.

Table 4: Suppressed Hospital Records

Age	ZIP_Code	Disease
45	12345	Flu
34	NA	Diabetes
28	12347	Hypertension
NA	12348	Asthma

Note: This table shows suppressed hospital records with certain suppressed values (replaced with NA) to protect privacy.

Anatomization and permutation serve different purposes compared to the aforementioned techniques. Rather than concealing QI details, these operations alter the relationship between QIs and SAs while preserving statistical properties. For example, suppose you observe a dataset as in Table 2 again with QIs being age and ZIP code, and a SA being medical condition. As seen in Tables 5 and 6, anatomization would split the dataset into two separate tables: Table 5 containing anonymized QIs (e.g. grouped age ranges and generalized ZIP code), and Table 6 containing the SAs (e.g. medical conditions). The separation provides for more control since different levels of access can be granted to different users, where researchers for example only receive the QIs table to analyze demographic trends. However, when both are used, the group ID column allows for group-based linking of the QIs and SAs without revealing direct association. On the other hand, permutation reshuffles sensitive values among data records. Again, consider the dataset in Table 2 where each record has a unique combination of QIs and SAs. Permutation would randomly reassign the SAs (e.g., medical conditions) among different records, disrupting the original data structure (see Table 7). This means that even if an attacker knows the QIs of a particular individual, they cannot accurately link it to the correct SA. However, the ability of permutation to prevent re-identification relies heavily on the level of permutation applied and must be selected with care to strike a balance between data privacy and utility.

Table 5: Anatomized Hospital Records: Quasi Identifiers

Group_ID	Age_Range	ZIP_Code_Range
1	40-49	12340-12349
1	40-49	12340-12349
2	30-39	12340-12349
2	30-39	12340-12349

Note: This table shows anatomized hospital records where QIs (age range and ZIP code range) are anonymized and separated from the SA.

Table 6: Anatomized Hospital Records: Sensitive Attributes

Group_ID	Disease
1	Flu
1	Hypertension
2	Diabetes
2	Asthma

Note: This table shows anatomized hospital records where the SA (Disease) is separated from the anonymized QIs.

Table 7: Permutated Hospital Records

Age	ZIP_Code	Disease
45	12345	Asthma
34	12346	Flu
28	12347	Diabetes
50	12348	Hypertension

Note: This table shows permutated hospital records with shuffled disease information to protect privacy.

Perturbation stands apart from other techniques as it results in data records that do not correspond to real-world individuals (Fung et al., 2010). Unlike anatomization and permutation, which preserve the relationship between QIs and SAs, perturbation modifies the actual values by introducing noise to dataset values while preserving statistical properties at an aggregate level. This noise can be drawn from a statistical distribution, such as a Gaussian distribution with a mean of zero. The amount of noise added is based on the standard deviation of the original data, ensuring that the overall properties remain similar. For instance, replacing geographic coordinates with randomly generated values obstructs pinpointing individuals' exact locations, which is a form of perturbation. In Table 8, new, unseen values can be observed in the age and ZIP code columns, demonstrating the application of perturbation.

Despite the benefits of these anonymization operations, they come with challenges. While some techniques ensure unchanged aggregate statistical properties, individual-level data utility may suffer. Thus, by applying these techniques you will settle for some data utility loss in order to gain more privacy protection. Determining an optimal balance requires assessing data usefulness and the level of privacy protection, highlighting the anonymization trade-off problem. Information metrics facilitate the answer to the trade-off problem and will be discussed for the specific privacy models conducted in this research in Section 3.

Table 8: Perturbated Hospital Records

Age	ZIP_Code	Disease
46	12344	Flu
33	12347	Diabetes
29	12346	Hypertension
51	12349	Asthma

Note: This table shows perturbated hospital records with slightly modified QI values to protect privacy.

2.2 Privacy Models

This section delves into the substantive literature review of privacy model, exploring the choices for incorporation into the research's privacy models. I aim to identify the most suitable existing model for k-anonymization and differential privacy within the scope of this study.

2.2.1 K-anonymization

The k-anonymity privacy model is widely adopted by researchers due to its cost-effectiveness and simplicity compared to alternative anonymity methods (Majeed and Lee, 2020). Academic literature offers numerous versions of the k-anonymity privacy model, each with distinct methodological approaches and differences.

Early works by Samarati (2001) and Sweeney (2002) introduced k-anonymity models that aim to achieve the optimal privacy threshold while minimizing the impact on data utility and computational resources. These models use complex algorithms to find the best possible generalization and suppression strategies evaluating every combination of anonymized attributes. However, achieving optimal k-anonymity is computationally expensive and is classified as an NP-hard problem (Fung et al., 2010). This classification implies that as the dataset size increases, the time required to find an optimal solution escalates rapidly, making it impractical to find a solution in polynomial time. To address these challenges, LeFevre et al. (2005) proposed Incognito, a collection of optimal generalization algorithms using bottom-up approaches. This method starts with the most generalized form and refines it if the refined version is more useful while still protecting privacy and, in such manner, systematically searches the space of all possible generalizations. Despite these improvements, the computational demands remain high due to implementation on the full attribute space, limiting the practicality for large datasets.

In response to the computational challenges of optimal models, researchers also developed minimal anonymous models that meet basic privacy requirements with more manageable computational demands. These models do not seek a perfect solution and may accept more aggressive generalization if the k-anonymity is met. One of the earliest minimal anonymous k-anonymity models was introduced by Hundepool and Willenborg (1996) with the μ-Argus algorithm. The μ-Argus approach quickly generalizes data to meet basic privacy requirements but does not consider more detailed groupings.

While effective for small datasets, it struggled with scalability and could handle only a limited number of attributes. The combinatorial nature of anonymizing multiple attributes and the increased complexity of handling dependencies between attributes led to significant computational challenges. Hence, Sweeney (1998) introduced Datafly, a heuristic-based algorithm that employed a greedy approach. Datafly generalizes data by starting with small generalizations and incrementally increasing the level of generalization until the k-anonymity criterion is met. This step-by-step technique is more efficient and can handle larger datasets, although its greedy nature sometimes led to excessive generalization, resulting in a loss of data utility.

LeFevre et al. (2006) developed the Mondrian Multidimensional k-anonymization algorithm, which improved upon Datafly by introducing a multidimensional partitioning approach. Mondrian partitions the data into regions that are generalized independently, allowing for more flexible and efficient anonymization. This approach utilizes a relaxed constraint, dynamically adjusting its partitioning strategy based on the data's multidimensional structure rather than strictly adhering to a single path of generalization as Datafly does. Consequently, this method strikes a balance between privacy and utility by preserving the multidimensional structure of the data while ensuring k-anonymity and is therefore argued to be a combination of both the optimal and minimal approach. Mondrian's ability to handle large datasets and its superior performance in terms of data utility and computational efficiency have made it a widely used and well-regarded k-anonymization method in the literature.

Given the need to balance data utility and computational efficiency while ensuring robust privacy protection, this study employs the Multidimensional Mondrian k-anonymization model by LeFevre et al. (2006). Its methodological advantages, particularly in handling large datasets, make it the most suitable choice for the analysis in this research.

2.2.2 Differential Privacy

Although the Mondrian Multidimensional k-anonymity privacy model outperforms many others and works relatively well on large data sets, privacy attacks remain a concerning challenge within k-anonymity. Especially when a data set is not only large but scarce too, identity disclosure attacks as well as attribute disclosure attacks can still endanger the privacy of the individuals in the data set. Therefore, I also turn to implementations of the differential privacy framework that have been increasingly analyzed the past few years and explain why the IPF method (Nowok, 2016) with the extension of differential privacy (Raab, 2022) is the most suitable choice for this study.

Dwork (2006) suggested the strong theoretical framework differential privacy which guaranteed the privacy of individuals in a dataset by making sure the presence or absence of the individual has no effect on the outcome distribution. The initial implementation of differential privacy by Dwork (2006) proposed to use the Laplace mechanism. This method adds noise drawn from a Laplace distribution directly to individual data points and thereby effectively obfuscates the individual data points. The amount of noise added is proportional to the distribution of the original data and controlled by the privacy budget ϵ. While this method is relatively simple to implement, applying

noise directly to individual data points can result in significant noise, particularly in datasets with correlated attributes as the added noise can amplify these correlations, leading to distorted results. This approach can therefore significantly degrade data utility (Zhu et al, 2017).

As research progressed, more sophisticated differential privacy mechanisms were developed to address the limitations of the Laplace mechanism. These include noise addition drawn from different distributions. Researchers McSherry and Talwar (2007) investigated the Gaussian mechanism, which uses Gaussian noise, and the Exponential mechanism, which chooses resulting datasets using a scoring function that appoints higher probabilities to outcomes with greater utility. The mechanism uses a probability distribution over all the possible outputs, where each output's probability is exponentially proportional to its score, making it more likely to select high-utility outputs while still maintaining privacy. The Exponential mechanism is useful for non-numeric scoring based functions such as ranking, categorical choices, or decision-making processes. However, the technique can be computationally intensive due to the need to calculate a probability distribution over all possible outputs based on their utility scores. The Gaussian mechanism provides better accuracy compared to the Laplace mechanism when the data exhibits natural variance similar to a normal distribution. This approach results in more accurate outcomes under specific conditions but requires careful tuning of parameters such as variance, adding complexity to its implementation. Thus, although these mechanisms also have their limitations, the development of techniques enabled researchers to implement the most suitable approach according to their study.

As previously mentioned, data correlation can have significant implications when sharing complete datasets, as the correlations between quasi-identifiers can enable adversaries to infer sensitive information by making educated guesses based on a single known variable. To tackle this challenge, researchers have proposed various methods such as batch querying and synthetic data publishing to mitigate risks. Batch query solutions group similar data queries (i.e., data retrieval operations) together and answer them as a batch to minimize the impact of correlations on privacy. The determination of which queries are grouped into each batch is based on their similarity and relevance to one another. Xiao et al. (2010) suggested the Privelet method, which transforms the data by applying wavelet transforms that decomposes data into various frequency components (i.e., range-count queries) before adding noise and reconstructing data from these noisy coefficients. Although this method outperformed many prior methods and is relatively simple in terms of setting up and processing individual queries, this technique is limited to the predefined queries and therefore highly inflexible. On the other hand, synthetic data publishing involves generating new datasets that replicate the statistical properties of the original data to address correlation issues. My aim is to present a solution that is applicable to various types of analysis, considering that the specific purpose of the data recipient may be unknown within the scope of this thesis, so therefore developments in differentially private synthetic data publishing are highly beneficial for my research.

One of the first occurrences of differentially private synthetic data publishing was by the researchers Machanavajjhala et al. (2008). The researchers applied the Dirichlet-Multinomial model to count

data from the U.S. Census Bureau, employing a probabilistic model that accounts for the variability and uncertainty in the count data. However, they, along with numerous subsequent researchers, were unable to produce usable synthetic results. According to Schneider et al. (2018), the primary issue was the lack of covariates in the privacy models, which meant that the synthetic data did not accurately reflect the structure and dependencies of the original data. The inclusion of covariates by algorithms, such as the one proposed by Abowd et al. (2013), was identified as a key factor for success. However, while this approach succeeded for simple regression analysis, it did not perform well for multiple regression scenarios.

Academics concluded that relaxing the privacy constraints was necessary to address real-world cases, leading to the development of empirical differential privacy as outlined by Schneider and Abowd (2015). Their method adds noise to the data in a manner similar to differential privacy, but the noise addition is based on the likelihood of the posterior predictive distribution. This unbounded nature breaks the formal guarantees of differential privacy. Although these models offered practical solutions, critics pointed out that the lack of formal guarantees and the dependency on the choice of discretization rendered the measure not well-defined (Charest and Hou, 2017).

Fortunately, significant advancements in the field occurred with the introduction of Generative Adversarial Networks (GANs) and models fitted to data margins. GANs can integrate differential privacy by training two neural networks, the Generator and the Discriminator, against each other. The Generator produces synthetic data, and the Discriminator assesses its authenticity. By incorporating differential privacy into this process, GANs can produce synthetic data that accurately models complex data structures and relationships while ensuring privacy (Raab, 2022).

Nonetheless, marginal methods outperformed GANs by first ensuring each variable pair in the dataset is differentially private, fitting a model that only includes specific predefined important interactions between variables, and then adding noise to these interactions to maintain differential privacy. Nowok et al. (2016) introduced two high-performing and user-friendly marginal methods, one of which Raab (2022) successfully extended with the differential privacy framework, naming it the Iterative Proportional Fitting algorithm. This method adjusts cell counts in contingency tables to fit specified margins while adding noise to ensure privacy. This process maintains the overall distributional properties of the original data, generating synthetic data while ensuring differential privacy. Although the noise addition is drawn from the Laplace distribution, which may not be the best-performing technique, and the method is limited to categorical variables, it handles large and complex datasets with relatively low levels of noise, making it a highly suitable technique for this research.

In light of the developments in academic research on this topic and the scope of my thesis, I apply the IPF method with the differentially private extension proposed by Raab (2022) to assess its strengths and weaknesses and compare them to the Multidimensional Mondrian k-anonymization approach.

3 Methodology

In this section, the selected methodologies described in Section 2 will be elaborated upon. Although PPDP methods do not require extensive information about the dataset, some implementations of the methods exhibit preferences for certain data characteristics. Therefore, it is essential to first highlight key characteristics of the data and the data collection process. Subsequently, the details of the inter-methodology metrics (i.e., the prediction accuracies of the logistic regression model and the execution time) are explained. Finally, the mathematical foundation of Multidimensional Mondrian k-anonymization and the Iterative Proportional Fitting technique using differential privacy will be explained.

3.1 Data Collection

The dataset under consideration, gathered by the survey organization SSI in 2008 and previously used by researchers De Jong et al. (2012), contains 3,146 observations from individuals who participated in an online survey. As mentioned, these individuals represent a diverse cross-section of the global population, including nationalities from France, the United States of America, Brazil, and Singapore.

The selected variables are exclusively categorical in nature. As defined in the terminology section, the dataset includes one sensitive attribute and five quasi-identifiers. All direct identifiers have been entirely removed, and no non-sensitive attributes are included among the selected variables. The sensitive attribute is binary, indicating whether individuals have purchased prescription drugs without a prescription, which is considered illegal and highly confidential. To protect the confidentiality of the individuals, it is crucial to employ effective PPDP methods. Since both sensitive attributes and quasi-identifiers can potentially reveal the identities of individuals in a dataset, the methods will aim to preserve the privacy of both these kinds of variables. The quasi-identifiers refer to gender, education, social class, working status, and marital status. Among these, the variables $Gender$ and $MaritalStatus$ are binary, while the others are multilevel categories, primarily ordinal in nature. The dataset contains no missing values; however, some category levels include only a few observations. In Table 24 of the Appendix, the frequency and meaning of the various levels is stated This sparsity may pose challenges for the implementation of certain models, which will be addressed in the sections dedicated to each method.

3.2 Inter-methodology metrics

In addition to specific intra-methodology metrics for each PPDP technique, the inter-methodology metrics will assess the practicality of the methods to select the appropriate anonymization method for potential stakeholders. To compare the computational efficiency, the execution time for the anonymization step is recorded. The utility will be assessed by analyzing the changed prediction accuracy of the anonymized data compared to the original data using a logistic regression model. The logistic regression is constructed to predict the probability of the illegal prescription drug purchase (Y) as binary dependent variable based on several categorical quasi-identifiers as independent

variables. Each categorical variable has varying levels, and these levels are incorporated into the model using dummy variables where the first level is considered the reference category. The corresponding formula for the logistic regression model in this study can be expressed as:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) =$$

$$\beta_0 + \beta_{\text{Gender}} \cdot D_{\text{Gender}} + \sum_{j=1}^{6} \beta_{\text{Education}_j} \cdot D_{\text{Education}_j} + \sum_{j=1}^{5} \beta_{\text{SocialClass}_j} \cdot D_{\text{SocialClass}_j} +$$

$$\sum_{j=1}^{7} \beta_{\text{WorkingStatus}_j} \cdot D_{\text{WorkingStatus}_j} + \beta_{\text{MaritalStatus}} \cdot D_{\text{MaritalStatus}} + \beta_{\text{IllegalPurchase}} \cdot D_{\text{IllegalPurchase}}$$

where β_0 is the intercept, β_{Gender} is the coefficient for the dummy variable D_{Gender} (i.e., Female), $\beta_{\text{Education}_j}$ are the coefficients for the dummy variables $D_{\text{Education}_j}$ representing the levels of Education, $\beta_{\text{SocialClass}_j}$ are the coefficients for the dummy variables $D_{\text{SocialClass}_j}$ representing levels of Social Class, $\beta_{\text{WorkingStatus}_j}$ are the coefficients for the dummy variables $D_{\text{WorkingStatus}_j}$ representing levels of Work Status, $\beta_{\text{MaritalStatus}}$ is the coefficient for the dummy variable $D_{\text{MaritalStatus}}$ (i.e., Single), and $\beta_{\text{IllegalPurchase}}$ is the coefficient for the dummy variable $D_{\text{IllegalPurchase}}$ (i.e., No). For each categorical variable with L_i levels, $L_i - 1$ dummy variables are created. The intercept β_0 represents the log-odds of the outcome when all predictors are at their reference levels. The dummy variable coefficients represent the change in log-odds of the outcome for the corresponding category compared to the reference category.

First, the original data is used as input data to serve as a benchmark. Subsequently, both the Mondrian anonymized as well as the IPF anonymized data sets are used as input for the model. The logistic regression analyses are cross-validated 10-fold to obtain more reliable results. Furthermore, given the slightly skewed nature of the dependent variable (see Table 24 of the Appendix), metrics that account for this imbalance will be assessed, namely sensitivity, specificity, and balanced accuracy. Sensitivity refers to the correct classification of individuals who have illegally purchased prescription drugs, while specificity refers to the correct classification of individuals who have not. Balanced accuracy represents the equilibrium between these two metrics.

3.2.1 Verification statistical assumptions

Prior to describing the results of the logistic regression with the different inputs, it is essential to verify that the statistical assumptions of the model are satisfied. These assumptions include the independence of observations, a sufficient sample size, linearity of the logit, absence of multicollinearity, and absence of outliers.

First, the assumption of independence of observations is inherently satisfied in this study, as the data is collected through individual surveys completed independently by respondents. This independence

guarantees that there is no relationship between residuals.

As shown in Table 24 of the Appendix, the sample size of 3,146 observations is sufficient given the complexity of the model and the number of predictors, adhering to the rule of thumb of having at least 10 observations per independent variable.

Additionally, the assumption of linearity of the logit, where the relationship between the independent variables and the log-odds of the dependent variable is expected to be linear, is satisfied due to the categorical nature of the variables. It is crucial, however, that the categorical variables are properly coded and interpreted.

Multicollinearity occurs when two or more independent variables in the model exhibit high correlation, resulting in unreliable estimates of the regression coefficients. To detect potential multicollinearity, various tests such as Fisher's Exact Test, Cramer's V Test, and the Generalized Variance Inflation Factor (GVIF) were conducted on the original data.

Fisher's Exact Test assesses the independence between pairs of categorical variables, especially in situations where the frequency of pairs might be low. A p-value less than 0.05 indicates a significant association between the variables. As shown in Table 25 of the Appendix, several variable pairs exhibit p-values below 0.05, such as $Gender$ and $Education$ ($p = 0.0004998$) and $Gender$ and $WorkingStatus$ ($p = 0.0004998$), suggesting potential multicollinearity between these variables, warranting further investigation.

To measure the strength of association between two categorical variables, Cramer's V Test was examined. The values range between 0 and 1, with 0 signifying no association and 1 representing a perfect association. Observing the results in Table 26 of the Appendix, most values are low, indicating weak associations (0 to 0.1). However, moderate associations, with values between 0.1 and 0.3, were found for $Gender$ and $WorkingStatus$, as well as $WorkingStatus$ and $MaritalStatus$ The presence of these moderate associations necessitates additional testing to ensure reliable estimates.

The GVIF quantifies the extent of multicollinearity in a regression model and is adapted to handle multilevel categorical variables. A GVIF value of 1 indicates no multicollinearity, values between 1 and 5 suggest moderate multicollinearity, and values exceeding 5 indicate high multicollinearity within the model. Additionally, a transformed GVIF metric is evaluated, which scales the GVIF according to the degrees of freedom associated with each predictor. For this transformed metric, a value of 1 again signifies no multicollinearity, values between 1 and 2 suggest moderate multicollinearity, and values above 2 indicate high levels of multicollinearity. In this analysis, presented in Table 27, the GVIF values for all variables are close to 1, indicating no significant multicollinearity. Similarly, the transformed GVIF values remain very close to 1, suggesting an absence of multicollinearity. Hence, based on the analysis of various correlation tests, there is no significant evidence of (perfect) multicollinearity in the dataset, thereby satisfying this statistical assumption of the logistic regression model.

In addition to conducting multicollinearity tests, the presence of outliers is analyzed using Cook's distance and leverage plots. Outliers can disproportionately affect the model, so Cook's Distance is measured to identify influential points. The threshold for Cook's Distance is determined by an inverse proportion of the number of observations and the complexity of the model. This entails that in a larger dataset, a data point must have a more substantial impact to be considered influential. As the number of predictors increases, the threshold becomes larger, making the model more tolerant of influential data points. As presented in Figure 5 of the Appendix, most observations have values close to zero, indicating they are not overly influential. Although a few points exhibit higher values than the threshold, further analysis on these outliers is conducted. Consequently, the leverage measure visualization in Figure 6 of the Appendix indicates that most deviations of the independent variables from their mean lie within an acceptable range, although some high-leverage points are present. Based on these results, caution should be exercised as some data points may disproportionately influence the model. However, since the majority of observations remain close to the threshold, this is not necessarily problematic. Therefore, the final assumption of the logistic regression model is satisfied.

Given that all statistical assumptions are met, the logistic regression analysis can be performed.

3.3 Multidimensional Mondrian K-anonymization

The Multidimensional Mondrian approach by leFevre et al. (2006) broadly implements two steps, namely the partitioning and generalization step. I will now outline all the details to consider in this process.

3.3.1 Process

The Mondrian process initiates by recursively partitioning the dataset into smaller subsets until all partitions meet the k-anonymity criterion. This process is inherently greedy, selecting the optimal partition at each step without considering overall optimization, thereby enhancing computational speed. This characteristic allows the method to be applied to multidimensional data without excessive computational burden.

Unlike other methodologies that utilize single-dimensional partitioning, the Mondrian method employs a multidimensional partitioning approach. This approach considers multiple attributes simultaneously when determining how to partition the data, making it particularly effective for datasets with significant attribute correlations. During each iteration, the algorithm evaluates the importance of each quasi-identifier based on the number of unique values it contains. Mathematically, this is presented as follows: $I(QI) = \sum_{j=1}^{d} \text{unique}(QI_j)$, where $Unique(QI_j)$ is the number of unique values in quasi identifier QI_j. Attributes with a higher number of unique values have a greater potential for individual identification and are thus prioritized for partitioning. The algorithm seeks to maintain as much information as possible by selecting the attribute with the highest number of unique values.

The chosen attribute is then sorted, and a median split is performed, dividing the data into two subsets. The corresponding formula looks as follows:

$$\text{Median}(X) = \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2}$$

where X is the chosen attribute. After each split, the process verifies whether both resulting partitions independently satisfy the k-anonymity requirement, ensuring that each partition contains at least k indistinguishable records based on the QIs. The mathematical representation of the criterion looks as follows: $|E_i| \geq k$, where E_i represents the equivalence class. This study employs both strict partitioning as well as relaxed partitioning. Strict partitioning treats each partition as an isolated subset where the k-anonymity requirement must be independently satisfied, while relaxed partitioning allows for overlapping regions. The strict partitioning technique ensures robust privacy guarantees but may lead to larger equivalence classes and therefore potentially higher information loss. On the other hand, relaxed partitioning blurs distinct boundaries by allowing for overlapping regions, but the increased flexibility can also result in better data utility. To fully comprehend the difference between the two techniques, a simplified example is shown in Table 9. For strict partitioning each group is strictly defined by both age and city with no overlap between groups, while for relaxed partitioning the groups are defined by age only, allowing for overlap between people from different cities within the same age range. With that logic in mind, if both partitions meet the criterion, the process continues by calculating the importance measure, selecting the most important attribute, and performing another median split. If a partition cannot be split further without violating the k-anonymity requirement (i.e., a resulting partition would have fewer than k records), the partitioning process stops for that partition, and the algorithm proceeds to the next step: recoding.

Table 9: Partitioning of individuals based on age and city

Person	Age	City	Strict.Partitioning	Relaxed.Partitioning
Alice	25	New York	Group 1: 20-30, NY	Group 1: 20-30
Bob	27	New York	Group 1: 20-30, NY	Group 1: 20-30
Charlie	35	Boston	Group 2: 30-40, Boston	Group 2: 30-40
Dana	29	Boston	Group 3: 20-30, Boston	Group 1: 20-30
Eve	22	New York	Group 4: 20-30, NY	Group 1: 20-30
Frank	42	Boston	Group 5: 40-50, Boston	Group 3: 40-50

Note: This table presents the strict and relaxed partitioning techniques. The strict partitioning column shows strictly defined non-overlapping groups by both age and city. The relaxed partitioning column is defined by age and shows overlapping groups for different cities.

Recoding entails the generalization or suppression of values or attributes to protect the identity of individuals while ensuring data utility. This can be implemented at either a global or local level. Local recoding modifies values within individual records, tailoring adjustments to the specific needs

of each record. On the other hand, global recoding defines a fixed set of generalized values for each attribute, uniformly applied across the entire dataset. For illustrative purposes, Table 10 presents local recoding of the city attribute by labelling in finer distinctions such as "NY Metro" and "Boston Area", while global recoding applies a uniform rule, labelling the cities into the broader region "Northeast USA". This study adopts the global recoding approach, ensuring consistent recoding throughout the dataset.

Table 10: Recoding of individuals based on age and city

Person	Age	City	Local.Recoding	Global.Recoding
Alice	25	New York	Age: 20-30, City: NY Metro	Age: 20-30, City: Northeast USA
Bob	27	New York	Age: 20-30, City: NY Metro	Age: 20-30, City: Northeast USA
Charlie	35	Boston	Age: 30-40, City: Boston Area	Age: 30-40, City: Northeast USA
Dana	29	Boston	Age: 20-30, City: Boston Area	Age: 20-30, City: Northeast USA
Eve	22	New York	Age: 20-30, City: NY Metro	Age: 20-30, City: Northeast USA
Frank	42	Boston	Age: 40-50, City: Boston Area	Age: 40-50, City: Northeast USA

Note: This table presents the local and global recoding of individuals based on their original age and city. The difference in recoding is observed in the city variable where local recoding makes finer distinctions and global recoding applies a uniform rule.

After partitioning the dataset, global recoding is applied. For each QI within a partition, actual values are replaced with a range spanning from the minimum to the maximum value found in that partition. The mathematical formula which corresponds to this is: $[\min(QI_i), \max(QI_i)]$. This step effectively anonymizes the data by removing precise value details and replacing them with a generalized interval that retains data truthfulness while protecting individual identities. Additionally, this recoding ensures data usability by providing sufficient detail to understand value distributions without disclosing exact information.

3.3.2 Evaluation Metrics

To assess the performance of the method, specific inter-methodology metrics are defined for evaluating the protection of individuals, data utility, and computational efficiency.

For protection purposes, it is most straightforward to verify whether the data satisfies the criterion of equivalence classes containing at least k records. To establish stronger privacy protection boundaries, the k-value can be increased. However, this may reduce data utility, necessitating a metric to further analyze data utility. To evaluate multiple levels of privacy protection, k-values 2, 5, 10, 15, 20 are used.

For both protection and data utility purposes, the Discernibility Metric (DM) is calculated similar

as proposed by LeFevre et al. (2017) using the formula:

$$C_{DM} = \sum_{E \in \text{EquivClasses}} |E|^2$$

where E represents an equivalence class and $|E|$ determines the size (number of records) of equivalence class E. The quadratic nature of the formula signals that larger equivalence classes have a disproportionately higher impact on the DM. An increased DM value generally implies enhanced privacy protection at the expense of data utility, as the anonymization process has deemed more records identical.

3.3.3 Challenges

As previously noted, the Mondrian algorithm can handle both numeric and categorical data. However, the implementation outlined by LeFevre et al. (2006) and employed in this study is limited in that it converts categories into numerical values before applying the algorithm. Given that the selected variables are primarily ordinal or binary, this limitation is not overly problematic, although it does result in some additional information loss.

Moreover, dataset sparsity can complicate partitioning, or potentially even prevent the algorithm from anonymizing the data and also satisfying the k-equivalence class requirement.

3.4 Iterative Proportional Fitting with Differential Privacy

While numerous methods exist to create differentially private datasets, generating synthetic versions of the original data is most desirable for this thesis. To create synthetic data using Iterative Proportional Fitting with differential privacy, several critical steps are necessary, each designed to balance the trade-off between data privacy and utility (Raab, 2022). This section outlines these steps.

3.4.1 Process

The IPF method supports only categorical variables; therefore, ensuring the dataset's suitability during preprocessing is crucial. Continuous variables can be utilized by discretizing them into categorical bins. Furthermore, the method does not function properly with missing values, as these will introduce biases and inaccuracies. Consequently, missing values should be either imputed or excluded. Fortunately, the dataset in this thesis contains neither continuous variables nor missing values, thus obviating the need for preprocessing steps.

To ensure differential privacy, a constraint is set using the parameter ϵ. ϵ, the privacy budget, controls the balance between privacy and utility, representing the cumulative sensitivity of individual data entries within the dataset. The core principle of the ϵ -differential privacy framework is encapsulated in the following mathematical expression:

$$\Pr[\mathcal{A}(D) \in S] \leq e^{\epsilon} \times \Pr[\mathcal{A}(D') \in S]$$

where $\mathcal{A}$ represents the randomized algorithm applied to the dataset; D and D' are two datasets differing by at most one individual's data; S is a set representing any possible subset of outputs from the algorithm; and e^{ϵ} is an exponential factor based on ϵ which controls the degree of privacy by bounding how much more likely any outcome is under one dataset versus the other. This framework ensures that the algorithm's output remains approximately the same, regardless of whether any individual record is included in the dataset, thereby safeguarding the privacy of individuals (Dwork, 2006). Lower ϵ values indicate a lower privacy budget to distribute and thus provide better privacy protection, even though it comes at the cost of reduced data utility. To assess the models' capabilities extensively, various values of ϵ are set and evaluated to find the preferred balance. In this study, 0.2, 0.5, 1, and 2 are the selected ϵ values, offering a spectrum of very strong privacy protections, default choices, and less privacy conservation options.

After data preparations and setting the privacy budget, cross-tabulations are created. As previously mentioned, cross-tabulations refer to the frequency distributions of combinations of variable values, which can be generated for all variables or for a predetermined selection of variable interactions, known as margins. To select margins the importance of interactions in the original data is observed. The default setting for creating cross-tabulations is based on the two-way interactions between all pairs of variables in the data. These pairwise interactions preserve key relationships without overly complicating the model. As Raab (2022) highlighted, the selected margins must be compatible with the subsequent data mining goals to ensure that the synthetic data remains useful. This means that the chosen margins should align with the intended analyses or tasks that the synthetic data will be used for, ensuring that key relationships and patterns relevant to those tasks are preserved. However, as discussed previously in Section 1.1 within the scope of this thesis it is unknown what the aim of the data miner is. Therefore, I opt for the generally effective approach for preserving important relationships by using pairwise interactions as margins.

To prevent cross-tabulation cells from having zero counts and ensure statistical integrity by maintaining a minimal presence for every possible combination of variables, a value is added to each cell in every cross-tabulation. This value is determined by setting $priorn > 0$ and dividing it by the number of cells in the frequency table (n_{cells}). Parameter *priorn* must be positive as a zero value would leave the frequency tables unchanged, and a negative value could reduce counts to negative values, which is meaningless for frequencies. To evaluate the outcomes of applying various levels of smoothing, various *priorn* values are tested. Typically, a small value is selected to cause only minimal distortion. Therefore, 0.1, 0.5 and 1 are chosen as well as 0 for *priorn* to compare the smoothing parameters with the baseline. For robustness reasons, the results of testing the selected values of the *priorn* parameter with the selected values of the ϵ parameter are averaged over 15 syntheses.

Next, noise drawn from the Laplace distribution is added to the counts in each cell of the cross-tabulation to introduce controlled randomness into the data, thereby protecting the privacy of individuals. Laplace noise is commonly used in differential privacy due to the mathematical properties of its Probability Density Function (PDF). The Laplace distribution PDF is presented by the following formula:

$$f(x|m,b) = \frac{1}{2b} \exp\left(-\frac{|x-m|}{b}\right)$$

where x is the variable, m is the mean of the distribution, and b is the scale parameter. The distribution is symmetric and has a sharp peak at its mean. Consequently, most noise values will be close to zero, minimally biasing the data, but some values will be farther away depending on the scale parameter b. The scale parameter for the Laplace distribution is calculated by the following formula: $b = \frac{M}{\epsilon}$, where M is the number of margins and ϵ is the privacy budget. By definition of the scale parameter, the likelihood of larger added values increases, resulting in more perturbed values, as the privacy budget decreases.

After adding Laplace noise to the cells in the frequency tables, some cells may contain negative counts due to the distribution's symmetry and zero mean. These negative counts are set to zero, as negative frequencies are illogical. Additionally, it is crucial to then scale the counts to form a valid probability distribution summing to unity, accurately representing the original distribution.

With the noisy margins created, the IPF model is applied. The algorithm iteratively adjusts the initial margins (i.e., the scaled original frequency, which serve as the prior probabilities) until they are consistent with the noisy marginal distributions including the Laplace noise and thus compatible. The updated probabilities after taking into account the prior probabilities and the noisy margins, are called the posterior probabilities. During each iteration, the algorithm first calculates the discrepancy between the noisy margins and the initial margins. Based on the size of this discrepancy, the algorithm scales the probabilities of each initial margin and thereby bringing them closer to the noisy margins while maintaining their relative proportions. This process repeats until the margins gradually converge to a stable solution ensuring privacy while also preserving the structure and relationships within the original data as accurately as possible.

After the IPF adjustments, each cell in the contingency table has a fitted probability. The synthetic data is generated by sampling from the multinomial distribution using these fitted probabilities. The multinomial distribution parameters are derived from the posterior probabilities for each cell, as determined by the IPF method. The number of synthetic data points generated is equal to the number of individuals contained in the dataset. The log-linear fit is used to model relationships between multiple categorical variables in the contingency table, ensuring that the synthetic data preserves these relationships.

The variables are grouped and synthesized in a specific sequence, known as the visit sequence, to

ensure that relationships between them are accurately reflected in the synthetic data. For example, after generating synthetic data for gender and education, the synthetic values of these variables are used to conditionally generate synthetic data for marital status, thereby preserving the relationships between these variables.

3.4.2 Evaluation Metrics

In evaluating the performance of differentially private synthetic data, researchers employ various metrics to determine utility and privacy protection. This thesis will assess metrics similar to Raab (2022) by comparing metrics for both a synthetic data set incorporating differential privacy and a synthetic data set without differential privacy constraints.

To assess privacy protection, the replicated uniques are determined. Replicated uniques are defined as the proportion of unique observations in the synthetic data that remain unique in the original data as well. The formula is as follows:

$$ru = \left(\frac{\sum_{i=1}^{k} (s_i = 1 \,\text{and}\, y_i = 1)}{k} \right) \times 100$$

where s_i and y_i are the counts in the synthetic and original data respectively for cell i in the cross-tabulation and k is the number of cells in the cross-tabulations. A higher percentage of replicated uniques indicates a higher risk of compromising the identity of individuals in the data set, as it reflects greater similarity between the synthetic and original datasets. Therefore, a lower percentage is generally desirable, though it may come at the cost of information loss. Nonetheless, this value depends on the uniqueness of the original data. Hence, to scale the value by the uniqueness of the original value, I also determine the following privacy metric: $(ru \,\text{as a}\, \% \,\text{of}\, p1) = \left(\frac{ru}{p1}\right) \times 100$, where $p1$ is defined as the percentage of unique records in the original data with the formula $p1 = \left(\frac{\sum_{i=1}^{k} (y_i = 1)}{N} \right) \times 100$. Here, again y_i is the number of counts in the original data for cell i in the cross-tabulation and N is the total number of records in the original data.

To measure the utility of differentially private synthetic data, the propensity score Mean Square Error (pMSE) is introduced. As Raab (2022) stated, pMSE assesses utility by first computing propensity scores for each observation in both the original and synthetic datasets. This involves estimating the probability that an observation belongs to the synthetic dataset versus the original dataset based on the covariates or variables available in the dataset. Then, the squared differences between the true propensity scores in the original data and those estimated from the synthetic data are computed for each observation and summed across all observations. Finally, the average of the sum of squared differences is taken to obtain pMSE. The corresponding formula for pMSE is: $\text{pMSE} = \frac{1}{N} \sum_{i=1}^{N} (\hat{p}_i - p_i)^2$ where $\hat{p}_i$ is the estimated propensity score for observation i in the synthetic data, p_i is the true propensity score for observation i in the original data, and N is the total number of observations. A lower pMSE value indicates a smaller discrepancy between the true and estimated propensity scores, signifying better utility. The pMSE values for both synthetic

and differentially private synthetic data are computed to assess the differential privacy constraint's impact.

3.4.3 Challenges

Generating differentially private synthetic datasets with IPF can present challenges in certain situations. For example, high-dimensional datasets may lead to sparse contingency tables and unreliable estimates of joint variable distributions during the IPF process, affecting the quality of the generated synthetic data.

Additionally, significantly large datasets might cause computational complexity to grow exponentially due to the increased number of contingency tables. Therefore, the original dataset must have a manageable number of variables to ensure the model executes properly and converges to a compatible dataset within a reasonable timeframe.

Furthermore, the purpose of the data mining task after publishing the differentially private synthetic data is uncertain. The data may be incompatible with the analysis performed at the data mining stage. While pairwise interactions generally perform well, expert domain knowledge of the attributes and data mining tasks could optimize the model and thus the generated data.

Lastly, the probabilities calculated during the IPF process may not sum to exactly 100 percent due to rounding errors or numerical instability. While small deviations may not significantly impact the synthetic data's overall utility, larger discrepancies can affect the dataset's accuracy of the generated dataset.

4 Results

In this section, I will analyze the outcomes of the selected and implemented methodologies. First, the results of the Multidimensional Mondrian K-anonymization, as introduced by LeFevre et al. (2006), are presented and explained. Second, the results of the Iterative Proportional Fitting (IPF) method are displayed, both with and without the incorporation of the differential privacy framework, as first presented by Raab (2022). Third, based on the results of the conducted anonymization algorithms, I will determine the optimal values for the hyperparameters and use them to create two anonymized datasets—one for each method. These anonymized datasets will then serve as input for the logistic regression analysis to compare the inter-methodology metrics.

4.1 Mondrian k-anonymization

The Multidimensional Mondrian K-anonymization algorithm is characterized by one critical parameter, k, which determines the level of privacy protection. Consequently, the algorithm is evaluated by assessing various k-values. As for example illustrated in Table 11, the k-values analyzed include 2, 5, 10, 15, and 20. Higher k-values correspond to stronger privacy protection; however, this increased privacy invariably comes at the cost of data utility, as has been repeatedly discussed.

4.1.1 Strict partitioning

Initially, the various k-values were assessed for the Mondrian k-anonymization with strict partitioning. As shown in Table 11, the k-anonymity criterion was not satisfied for any of the strictly partitioned anonymized datasets. Further analysis revealed that many partitions generated by the method described in Section 3.2.1 contained unique records, rendering them unsuitable. This suggests that the dataset's high uniqueness presents a challenge for this approach. To provide additional context, summary statistics are presented in Table 12, illustrating that 20.47 percent of the original dataset comprises unique records. This significant proportion leads to partitions being overly granular, resulting in groups with fewer than k indistinguishable records.

Table 11: Results of Strict Partitioned Mondrian Anonymization

k	DM	k_Anonymity
2	120476	FALSE
5	101178	FALSE
10	96948	FALSE
15	111930	FALSE
20	111930	FALSE

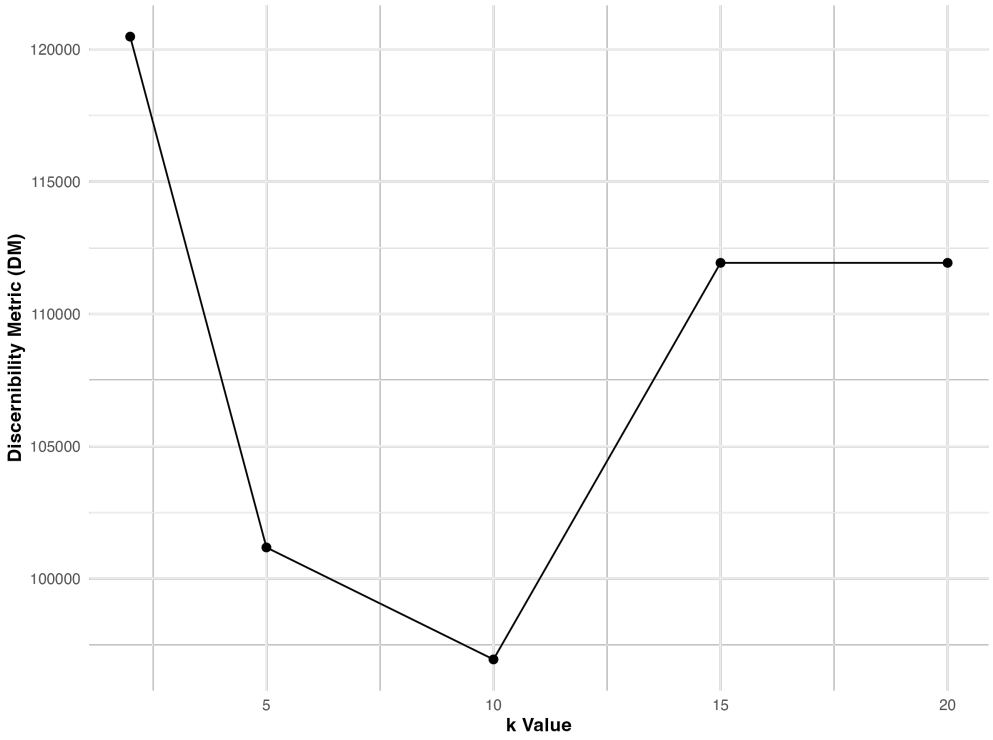
Note: This table presents the results of strict partitioned Mondrian anonymization. The k column indicates the k-values used in the anonymization process. The DM column shows the Discernibility Metric for each k-value, and the k Anonymity column indicates whether k-anonymity was achieved (TRUE) or not (FALSE).

Table 12: Summary statistics original data

Metric	Value
N	3146.00
n_cells	2688.00
p0	76.04
p1	20.47

Note: This table summarizes the key metrics from the original dataset. The Metric column includes various statistical measures, and the Value column provides their respective values. The various statistical measures represent respectively the number of observations, the number of cells in the cross-tabulations, the percentage of empty cells in the contingency tables, and the percentage of uniques in the original data.

Figure 2: Privacy-utility Trade-off for Strict Partitioning



Note. This figure illustrates the relationship between the Discernibility Metric (DM) and various k values under strict partitioning conditions for the Multidimensional Mondrian k-anonymization method.

Despite the failure to meet the k-anonymity criterion, where at least one equivalence class is smaller than k, the DM can still be calculated to obtain an indication of the overall anonymity cost of the dataset. As indicated in Table 11 and Figure 2, the DM value generally increases with higher k-values. This aligns with the expectation that stronger privacy protection typically reduces data utility. However, extreme caution should be exercised when interpreting the absolute DM values, as the algorithm failed to apply sufficient privacy protections to meet the threshold. Consequently, it is imperative to evaluate an alternative technique that may fulfill the k-anonymity criterion,

specifically the relaxed partitioning Mondrian technique.

4.1.2 Relaxed partitioning

The relaxed partitioning constraints of the Mondrian algorithm, which permit overlapping regions, aim to maintain better data utility while still respecting privacy requirements. Implementing this approach on the dataset satisfies the k-anonymity criteria for various k-values (see Table 13). This indicates that each record in any given partition has at least $k-1$ indistinguishable counterparts, effectively anonymizing the data.

Table 13: Results of Relaxed Partitioned Mondrian Anonymization

k	DM	k_Anonymity
2	35106	TRUE
5	102962	TRUE
10	287330	TRUE
15	359592	TRUE
20	359592	TRUE

Note: This table presents the results of the relaxed partitioned Mondrian anonymization process. The k column represents different k-anonymity levels, the DM column shows the discernibility metric, and the k Anonymity column indicates whether k-anonymity was achieved (TRUE) or not (FALSE).

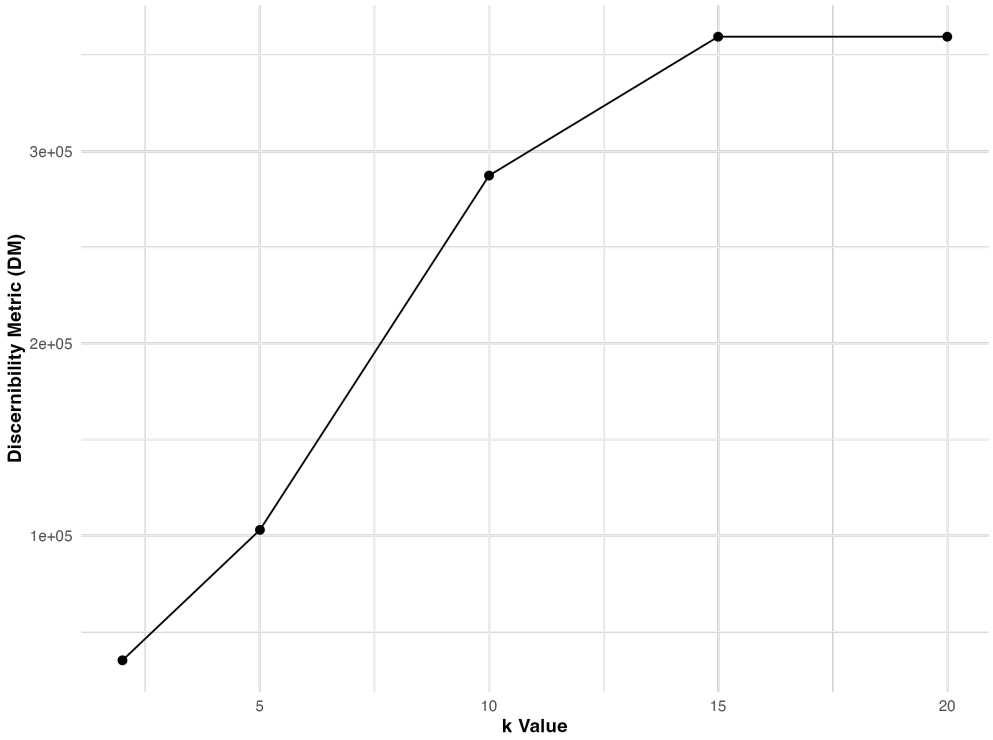
Regarding the DM value, it generally increases as k increases, peaking at $k = 15$ (see Figure 3). This trend is typical, as higher DM values indicate more records being deemed identical due to stronger generalization operations, enhancing privacy at the expense of data utility.

Moreover, variations in the steepness of the DM curve are observed. The steeper increase between $k = 5$ and $k = 10$ typically suggests that achieving $k = 10$ anonymity necessitated substantial data generalization. This can occur when the QIs within existing partitions at $k = 5$ exhibit high variability or numerous outliers. Consequently, even a small increase in k may disproportionately impact certain data subsets, requiring significantly more generalization to ensure privacy compliance. As illustrated in Tables 28 and 29 of the Appendix, previews of the anonymized datasets for the *SocialClass* attribute at $k = 5$ and $k = 10$ show that stronger privacy requirements led to generalizing from $[2-4]$ to $[2-6]$. This substantial generalization step likely affects other partitions similarly.

Interestingly, the DM curve flattens after $k = 15$, indicating that further increases in k do not require significant additional generalizations to maintain k-anonymity. This phenomenon can be understood by examining both the theoretical implications of the DM curve and the actual data transformations at $k = 15$ and $k = 20$ shown in the provided examples (Tables 30 and 31 in the Appendix). Theoretically, this may occur if most QIs within the dataset are already grouped into broad categories satisfying the k-anonymity criterion for higher k-values. Thus, data publishers can select a k-value between 15 and 20 without significant losses in data usability. Practically, the lack

of change in range categories between $k = 15$ and $k = 20$ illustrates why the DM may plateau; since the ranges are not expanding further, the "penalty" measured by the DM for anonymization does not increase.

Figure 3: Privacy-utility Trade-off for Relaxed Partitioning



Note. This figure illustrates the relationship between the Discernibility Metric (DM) and various k values under relaxed partitioning conditions for the Multidimensional Mondrian k-anonymization method.

4.2 IPF with Differential Privacy

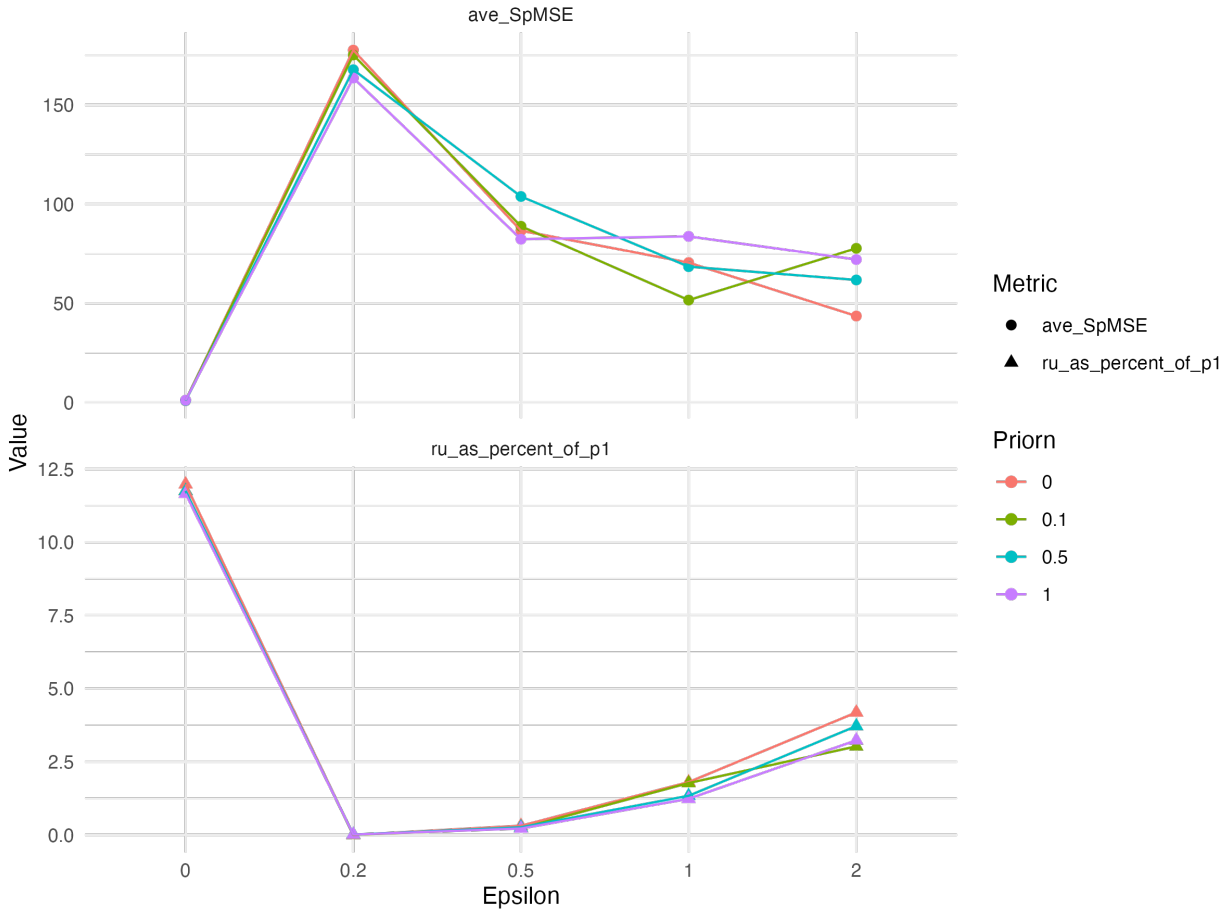
In addition to the Mondrian approach, this section analyzes and interprets the results of the IPF algorithm, particularly its extension incorporating the differential privacy framework. This technique allows for the adjustment of two key parameters: ϵ and *priorn*, based on the preferences of the data publisher. As previously mentioned, increasing ϵ values results in lower levels of privacy protection. The evaluation considers ϵ values of 0.2, 0.5, 1, and 2, comparing these results both within the synthetic datasets and against those generated without differential privacy (i.e., $\epsilon = 0$). The impact of the *priorn* parameter, assessed at values 0, 0.1, 0.5, and 1, is also examined to understand its role in data smoothing and maintaining statistical integrity.

4.2.1 Non-DP synthesis

First, the outcomes of generating synthetic datasets without differential privacy constraints are evaluated by analyzing the results for $priorn = 0$. As shown in Figure 4 (and more precisely in Table

14), the replicated uniques metric is relatively high, with a value of 2.45%. This means that 2.45% of the records that are unique in both the synthetic and original datasets can potentially be linked back to unique individuals in the original data. Although this percentage seems low, it is significant because these unique records could potentially be reidentified. Intuitively, non-differentially private synthetic datasets result in higher reidentification risks since the algorithm does not add differential privacy noise to the data. The differences relative to the original data arise only from the random generation of records based on probabilistic sampling from the contingency tables. The IPF process ensures that the synthetic data has the same marginal distributions as the original data, although individual records might differ.

Figure 4: Utility and Reidentification Risk for IPF method



Note. The upper plot of the figure illustrates the utility by quantifying the difference in distribution of the anonymized versus the original dataset. The bottom plot of the figure illustrates the reidentification risk by determining the replicated uniques (ru) as a percentage of the uniqueness of the original dataset (p1). Both plots present the metrics for different privacy budgets (ϵ) and priors (*priorn*) and are derived from the IPF algorithm averaged over 15 syntheses.

Table 14: Results for Epsilon = 0

Priorn	RU	RU as % of p1	Average standardized pMSE
0.0	2.45	11.98	0.90
0.1	2.41	11.75	1.03
0.5	2.41	11.75	0.94
1.0	2.39	11.66	1.13

Note: Note. This table presents the results for varying priors for non-DP syntheses. The columns show the prior values, reidentification risk (RU), reidentification risk as a percentage of uniqueness (p1), and average standardized mean square error (S pMSE). The results are averaged over 15 syntheses.

Despite the high reidentification risk in non-DP synthesis, the average standardized pMSE value of 0.90 indicates high data utility (see Table 14). This metric reflects the dissimilarity of the synthetic distribution compared to the original distribution, and a low value indicates a high level of resemblance. The lack of noise addition preserves data utility, making potential data mining tasks performed on the synthetic data yield results very similar to those obtained from the original data.

Next, the non-DP synthetic dataset is analyzed while varying the *priorn* parameter. The risk of reidentification remains consistent across different *priorn* values, slightly decreasing as *priorn* increases. For example, the replicated uniques metric decreases from 2.45% to 2.39% when *priorn* increases from 0 to 1. These minor changes suggest that increasing *priorn* slightly enhances data privacy by reducing the reidentification risk. This is because *priorn* prevents cells in contingency tables from having zero counts, distributing records more evenly across the contingency table. However, the impact is minimal, indicating that *priorn* has a limited effect on privacy in the non-DP setting.

Conversely, the utility metric fluctuates more noticeably with changes in *priorn*. The average standardized pMSE value increases from 0.90 to 1.13 when *priorn* increases from 0 to 1. This indicates that higher *priorn* values lead to increased utility loss. Although preventing zero counts aims to maintain statistical integrity, in this case, it introduces minor distortions that lead to a small loss in utility.

4.2.2 DP syntheses

To assess higher levels of privacy protection, the outcomes of the DP-synthetic datasets shown in Tables 15 to 18 and Figure 4 are analyzed. First, as observed in Table 15, the risk of reidentification is minimized for $\epsilon = 0.2$. The low privacy budget imposes strong privacy constraints on the generated synthetic datasets, fully preventing the reidentification of individuals and leading to 0 replicated unique records. However, achieving this high level of privacy protection comes at the cost of data utility. As seen in Figure 4, the dissimilarities are highest for $\epsilon = 0.2$, indicating the lowest data utility. This confirms the intuition of the trade-off between privacy and utility. Increasing the *priorn* parameter for $\epsilon = 0.2$ mitigates some of the utility loss caused by the noise, aligning with

the theory that it maintains the statistical properties of the distribution and enhances data utility. Thus, if the data publisher's primary interest is to preserve privacy, they should select the lowest value of ϵ (i.e., $\epsilon = 0.2$). To maximize utility for the chosen level of privacy, they should opt for a higher value of *priorn* (i.e., $priorn = 1.0$).

Notably, when the privacy budget increases from $\epsilon = 0.2$ to $\epsilon = 0.5$, there is a rapid decrease in the utility curve (see Figure 4). This suggests that the utility of the data increases quickly as the anonymized distribution more closely resembles the distribution of the original dataset. This rapid improvement in utility with a relatively small increases in ϵ implies a higher sensitivity to changes in the privacy budget within this range. The application of slightly less noise makes the data significantly more useful for analysis, which is critical when balancing the need for privacy with the need for actionable data.

As the privacy protection constraints further decrease with increasing ϵ, the data utility of the resulting DP synthetic datasets significantly improves. As shown in Table 18, the lowest value of the average standardized pMSE metric is 43.57, indicating that the synthetic data retains more of the original data's structure and relationships, making it more suitable for analysis tasks by the data recipient. However, selecting the parameters that result in the highest utility also leads to the highest reidentification risk, again confirming the intuition of the trade-off problem. Nonetheless, the increase in replicated unique records is relatively low, suggesting only a moderate reidentification risk. Therefore, data providers who aim for optimal utility should select $\epsilon = 2$ and can do so without extreme losses in privacy protection. Contrary to the results for $\epsilon = 0.2$, utility diminishes as *priorn* increases. This could be because, in some instances, the addition of frequency counts destroys more utility due to distortions than it gains in maintaining statistical integrity. Thus, for $\epsilon = 2$, selecting $priorn = 0$ is most optimal.

After assessing the extremes in terms of utility and reidentification risk, a suitable choice to achieve a balance between reidentification risk and utility loss could be selecting parameters $\epsilon = 1$ and $priorn = 0.1$. The percentage of replicated uniques for this selection is only 0.36%, indicating a relatively low probability of an individual being unique in both the synthetic and original datasets. At the same time, the utility is relatively high (i.e., 51.60) compared to the highest value for $\epsilon = 2$. This ensures that the distribution of the anonymized dataset closely resembles that of the original dataset. To conclude, a preview of this anonymized dataset is shown in Table 32 of the Appendix.

In summary, the findings show that non-differential privacy syntheses present a high risk of reidentification of individuals while maintaining high levels of data utility, whereas differential privacy syntheses offer the flexibility to select a trade-off between data utility and privacy protection according to the data provider's preferences.

Table 15: Results for Epsilon = 0.2

Priorn	Average RU	Average RU as % of p1	Average S_pMSE
0.0	0	0.00	177.64
0.1	0	0.00	175.18
0.5	0	0.01	167.76
1.0	0	0.01	163.58

Note: This table shows results for varying priors with a fixed epsilon, including prior values, average reidentification risk (RU), average reidentification risk as a percentage of uniqueness (p1), and standardized mean square error (S pMSE), averaged over 15 syntheses.

Table 16: Results for Epsilon = 0.5

Priorn	Average RU	Average RU as % of p1	Average S_pMSE
0.0	0.06	0.31	86.61
0.1	0.05	0.23	88.83
0.5	0.05	0.26	103.83
1.0	0.04	0.22	82.31

Note: This table shows results for varying priors with a fixed epsilon, including prior values, average reidentification risk (RU), average reidentification risk as a percentage of uniqueness (p1), and standardized mean square error (S pMSE), averaged over 15 syntheses.

Table 17: Results for Epsilon = 1

Priorn	Average RU	Average RU as % of p1	Average S_pMSE
0.0	0.37	1.80	70.50
0.1	0.36	1.77	51.60
0.5	0.27	1.34	68.47
1.0	0.25	1.23	83.72

Note: This table shows results for varying priors with a fixed epsilon, including prior values, average reidentification risk (RU), average reidentification risk as a percentage of uniqueness (p1), and standardized mean square error (S pMSE), averaged over 15 syntheses.

Table 18: Results for Epsilon = 2

Priorn	Average RU	Average RU as % of p1	Average S_pMSE
0.0	0.86	4.18	43.57
0.1	0.62	3.02	77.70
0.5	0.76	3.72	61.76
1.0	0.66	3.23	72.07

Note: This table shows results for varying priors with a fixed epsilon, including prior values, average reidentification risk (RU), average reidentification risk as a percentage of uniqueness (p1), and standardized mean square error (S pMSE), averaged over 15 syntheses.

4.3 Logistic Regression Analysis

After assessing the performance of the anonymization algorithms separately, this section compares the performance of both by analyzing the outcomes of a straightforward logistic regression using the resulting anonymized datasets as input. First, to establish a point of reference, the benchmark model is examined. The statistical assumptions of the original data and the benchmark model are already verified in Section 3.2.1.

4.3.1 Benchmark model

As shown in Table 19, the benchmark model exhibits a relatively low sensitivity of 43 percent, which measures the proportion of correctly identified positives. This low sensitivity suggests that the model misses many true cases, potentially leading to numerous false negatives. The specificity metric, or true negative rate, achieves a higher value of 63 percent, indicating a moderate ability to identify true negatives. However, this still leaves room for substantial improvement, as the balanced accuracy of 53 percent is only slightly better than random guessing in a balanced scenario. It is important to note that this study focuses not on creating the best data analysis model but on developing the best model for data publication. Consequently, these outcomes are used solely for comparison purposes, as a data publisher can easily adapt the input variables according to the preferences of data recipients, who can then optimize the data analysis model to improve prediction performance.

Table 19: Prediction performances

Metric	Benchmark	Mondrian	IPF
Sensitivity	0.43	0.50	0.65
Specificity	0.63	0.66	0.52
Balanced Accuracy	0.53	0.58	0.58

Note: This table summarizes the prediction performances from the 10-fold cross-validated logistic regression model with respectively the original dataset, the Mondrian anonymized dataset (with k=10 for relaxed partitioning), and the IPF dataset (with epsilon=1 and priorn=0.1) as input.

To compare whether the same coefficients are significant for the logistic regression models utilizing anonymized data, I first identify the significant variables in the cross-validated benchmark model, as shown in Table 20. The observed variables are compared to their reference category, which is the first level of each categorical variable in this analysis. Variables with a p-value smaller than 0.05 are considered significant, with $Gender2$ (i.e, Female), $WorkingStatus6$ (i.e., Retired), and $MaritalStatus2$ (i.e., Single) being noteworthy. $Gender2$ and $MaritalStatus2$ have a negative effect, while $WorkingStatus6$ has a positive effect on the log-odds of the dependent variable.

The significant benchmark model results indicate that being female, as opposed to male, decreases the log-odds of an illegal prescription drug purchase by 0.326. Similarly, having the status "Single" instead of "Married" decreases the log-odds of an illegal prescription drug purchase by 0.187. Conversely, a retired person, compared to a full-time worker, increases the log-odds of purchasing prescription drugs illegally by 0.413. However, it is important to note that the observed low balanced accuracy poses risks for interpreting these coefficient estimates, as the coefficients might reflect noise rather than true relationships. Moreover, this thesis primarily focuses on the potential differences in outcomes between the anonymized and original datasets. Therefore, the significance of the predictors in the anonymized datasets will be assessed first to determine if their significance persists.

Table 20: Summary of Cross-Validated Benchmark Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.1439267	0.9626095	-1.1883601	0.2346916
Gender2	-0.3262039	0.0851973	-3.8288040	0.0001288
Education2	0.6280944	1.0324337	0.6083630	0.5429468
Education3	1.4618656	0.9126093	1.6018526	0.1091882
Education4	1.1010995	0.8601645	1.2801034	0.2005088
Education5	0.9840745	0.8475870	1.1610306	0.2456295
Education6	1.1634402	0.8457403	1.3756471	0.1689309
Education7	0.9945399	0.8450420	1.1769118	0.2392307
SocialClass2	0.6428695	0.4682132	1.3730272	0.1697439
SocialClass3	0.4041739	0.4581465	0.8821937	0.3776720
SocialClass4	0.1272713	0.4645017	0.2739952	0.7840883
SocialClass5	0.5812884	0.4735196	1.2275910	0.2196005
SocialClass6	0.3718441	0.5437810	0.6838122	0.4940938
WorkingStatus2	-0.1409015	0.1466599	-0.9607366	0.3366846
WorkingStatus3	-0.5404357	0.2824584	-1.9133287	0.0557060
WorkingStatus4	-0.2191411	0.1927844	-1.1367164	0.2556569
WorkingStatus5	0.2066445	0.2111129	0.9788342	0.3276619
WorkingStatus6	0.4132260	0.1312121	3.1492975	0.0016366
WorkingStatus7	0.1253565	0.1734419	0.7227577	0.4698288
WorkingStatus8	-0.6575462	0.4167362	-1.5778474	0.1146007
MaritalStatus2	-0.1869903	0.0896351	-2.0861296	0.0369669

Note: This table summarizes the coefficients from the 10-fold cross-validated logistic regression model with the original dataset as input.

4.3.2 Mondrian model

To apply the logistic regression model to the Mondrian k-anonymized data, a specific k-value must be selected. Since the results will be compared to those of the IPF with differential privacy (DP) outcomes, it is essential to choose a similar trade-off between accuracy and privacy. Based on previous analyses, $k = 10$ was selected for the Mondrian method and $\epsilon = 1$ with $priorn = 0.1$ for the IPF with DP method.

Prior to performing the logistic regression analysis with the selected Mondrian data, it is essential to verify the statistical assumptions for the new dataset. Consistent with the reasoning of the benchmark model, the assumptions of independence of observations, sufficient sample size, and linearity of the logit are met. As shown in Table 33 of the Appendix, the transformed GVIF values, which account for the increased number of levels per independent variable, indicate that multicollinearity is absent from the Mondrian anonymized model, as all values are below the threshold of 2. An assessment of influential points using Cook's distance and leverage plots revealed results similar to those obtained from the original data analysis. Consequently, this final assumption is verified as well, allowing the logistic regression model to be conducted with the Mondrian data.

When observing the summary of the cross-validated Mondrian logistic regression model in Table 21, the different structure of the anonymized dataset is immediately notable. Due to relaxed partitioning, some original variable outcomes remain intact, while new variable outcomes are generated to represent generalized intervals, complying with anonymization requirements. Applying the logistic regression model to the Mondrian anonymized data necessitates transforming various possible outcomes of the variables into factor levels, resulting in a changed data structure. The reference category for every predictor observed in Table 21 is $Gender[1-2]$, $Education[1-7]$, $SocialClass[1-4]$, $WorkingStatus[1-2]$, and $MaritalStatus[1-2]$ respectively. This transformation diminishes data utility and should be considered when selecting an appropriate anonymization method.

The prediction performance has slightly improved with Mondrian k-anonymization as input data. As shown in Table 19, sensitivity increased to 50 percent, indicating a 50/50 chance of correctly identifying illegal prescription drug purchases. Specificity is even higher, with a 66 percent likelihood of correctly identifying individuals who did not illegally purchase prescription drugs, resulting in a balanced accuracy of 58 percent, which is slightly higher than the benchmark model. This improvement may be attributed to the relaxed partitioning nature, which preserves more detailed information in the QIs, providing more nuanced and informative predictor variables. Additionally, the enrichment of the feature set can help capture more complex relationships within the data, potentially improving predictive performance.

Table 21: Summary of Cross-Validated Mondrian Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.0551475	0.7174420	-1.4707076	0.1413702
Gender1	0.8093769	0.1727918	4.6841164	0.0000028
Gender2	0.4976939	0.1780488	2.7952673	0.0051857
'Education[2-7]'	0.5145025	0.3032249	1.6967686	0.0897404
'Education[3-7]'	0.3572931	0.2482152	1.4394491	0.1500233
'Education[4-6]'	-1.8643299	0.5656076	-3.2961544	0.0009802
'Education[4-7]'	0.4515398	0.2431070	1.8573711	0.0632584
'Education[5-7]'	0.2835497	0.2378404	1.1921845	0.2331889
'Education[6-7]'	0.8191970	0.2877572	2.8468339	0.0044156
'SocialClass[1-5]'	0.6128784	0.3129810	1.9581967	0.0502069
'SocialClass[1-6]'	0.5619170	0.4033945	1.3929714	0.1636284
'SocialClass[2-3]'	1.2026008	0.5520764	2.1783233	0.0293820
'SocialClass[2-4]'	0.2726212	0.2624082	1.0389204	0.2988417
'SocialClass[2-5]'	0.6978544	0.2474057	2.8206888	0.0047921
'SocialClass[2-6]'	0.5431128	0.2831210	1.9183059	0.0550722
'SocialClass[3-4]'	-0.5729136	0.3264510	-1.7549760	0.0792634
'SocialClass[3-5]'	0.7399565	0.2785260	2.6566876	0.0078913
'SocialClass[3-6]'	0.8571603	0.3179049	2.6962787	0.0070119
'WorkingStatus[2-3]'	-0.2765683	0.8892525	-0.3110121	0.7557914
'WorkingStatus[3-4]'	-0.2765683	0.8892525	-0.3110121	0.7557914
'WorkingStatus[4-5]'	-0.0304092	0.9540581	-0.0318735	0.9745729
'WorkingStatus[5-6]'	0.7143574	0.9057619	0.7886812	0.4302984
'WorkingStatus[6-7]'	-0.9284129	0.8938138	-1.0387095	0.2989399
WorkingStatus1	0.0712445	0.6475990	0.1100133	0.9123988
WorkingStatus2	0.1016973	0.6611763	0.1538126	0.8777575
WorkingStatus3	-0.4269271	0.7225387	-0.5908709	0.5546069
WorkingStatus4	-0.2845483	0.6814755	-0.4175474	0.6762781
WorkingStatus5	0.1216043	0.6872916	0.1769327	0.8595613
WorkingStatus6	0.6486176	0.6506958	0.9968062	0.3188586
WorkingStatus7	0.3213064	0.6870684	0.4676484	0.6400360
WorkingStatus8	-0.7774953	0.7663115	-1.0145944	0.3102992
MaritalStatus1	-0.1357101	0.2475557	-0.5482003	0.5835544

Note: This table summarizes the coefficients from the 10-fold cross-validated logistic regression model with a k=10 Mondrian anonymized dataset as input.

Lastly, the significance of the variables has changed substantially. At a 5 percent critical level, the variables $Gender1$ (i.e., Male), $Gender2$ (i.e., Female), $Education[6-7]$ (i.e., University and not disclosed), $SocialClass[2-3]$ (Upper middle class and Middle class), $SocialClass[2-5]$ (i.e., Upper middle class until Working class), $SocialClass[3-5]$ (i.e., Middle class until Working class), and $SocialClass[3-6]$ (i.e., Middle class until Lower class) show a positive significant effect, while

$Eduction[4-6]$ (i.e., Education up to age 18 until University) shows a negative significant effect on the log-odds of the outcome variable representing the purchase of prescription drugs without a prescription. The estimates of the coefficients changed as well and are not similar to the benchmark model.

4.3.3 IPF with DP model

As previously mentioned, the chosen parameters for the IPF technique with differential privacy constraints are $\epsilon = 1$ with $priorn = 0.1$. Again, the statistical assumptions must be verified for the logistic regression with the new IPF data. For this dataset, the assumptions of independence of observations, sufficient sample size, and linearity of the logit are satisfied. As indicated in Table 34 of the Appendix, the (transformed) GVIF values have not changed significantly and remain well below the threshold, confirming that multicollinearity is not an issue in this anonymized model. Additionally, an assessment of the differentially private data does not reveal any disproportionately influential points. Consequently, all the assumptions are met for this final model as well.

The results of implementing the parameter choices and using the output as input for the logistic regression model are presented in Tables 19 and 22. Starting with prediction performance analysis, sensitivity increased significantly compared to the benchmark model, reaching 65 percent. However, this improvement is balanced by a lower specificity metric of 52 percent, resulting in a balanced prediction accuracy of 58 percent, which is only slightly higher than the benchmark model's 53 percent.

Evaluating the summary statistics of the cross-validated IPF with DP logistic regression model in Table 22 reveals a data structure similar to that of the benchmark. However, the significance of the predictors has changed notably. While the benchmark model had only three significant independent variables, the IPF model has nine. The predictors with a positive significant effect on the log-odds of the outcome variable at a 5 percent significance level are $Education2$ (i.e., Education up to age 12), $Education3$ (i.e., Education up to age 14), $Education4$ (i.e., Education up to age 18), $Education5$ (i.e., Higher education), $Education6$ (i.e., University), $Education7$ (i.e., Not disclosed), and $WorkingStatus7$ (i.e., Housewife), while $Gender2$ (i.e., Female) and $MaritalStatus2$ (i.e., Single) have a significant negative effect.

Additionally, the spike in the estimate and standard error of $SocialClass6$ (i.e., Lower class) is particularly noticeable. This change is unlikely due to multicollinearity, as indicated by the GVIF values (see Table 34). Instead, it may result from the introduction of noise through the IPF with differential privacy process with as a result possible significant data sparsity in $SocialClass6$, which has only two observations. Both factors can increase the variability of the estimates. Therefore, the data utility of the IPF technique with DP appears suboptimal compared to the benchmark model.

Table 22: Summary of Cross-Validated IPF with DP Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.9853803	0.9097190	-1.0831700	0.2787330
Gender2	-1.1350152	0.0906422	-12.5219322	0.0000000
Education2	1.9067280	0.8604899	2.2158633	0.0267009
Education3	1.8340294	0.8419650	2.1782729	0.0293857
Education4	1.9188886	0.8386906	2.2879577	0.0221400
Education5	1.8495948	0.8396441	2.2028317	0.0276066
Education6	1.9692270	0.8434878	2.3346241	0.0195631
Education7	2.2267569	0.9063341	2.4568829	0.0140148
SocialClass2	-0.3179778	0.2845444	-1.1174982	0.2637813
SocialClass3	-0.2352018	0.2823605	-0.8329841	0.4048537
SocialClass4	-0.1527753	0.2926586	-0.5220258	0.6016524
SocialClass5	-0.1412281	0.3581372	-0.3943407	0.6933295
SocialClass6	11.5312251	196.9679245	0.0585437	0.9533156
WorkingStatus2	0.0618467	0.2528240	0.2446235	0.8067480
WorkingStatus3	0.1446858	0.2427609	0.5960010	0.5511745
WorkingStatus4	0.3211945	0.2431938	1.3207350	0.1865897
WorkingStatus5	0.3808643	0.2588892	1.4711478	0.1412511
WorkingStatus6	0.0217083	0.2998822	0.0723893	0.9422921
WorkingStatus7	1.6500401	0.5611167	2.9406360	0.0032754
WorkingStatus8	0.4484687	1.0299981	0.4354073	0.6632668
MaritalStatus2	-0.3072133	0.1097107	-2.8002121	0.0051069

Note: This table summarizes the coefficients from the 10-fold cross-validated logistic regression model with an IPF anonymized dataset (with epsilon=1 and priorn=0.1) as input.

4.3.4 Comparison

Having compared the anonymized logistic regression models to the benchmark, it is now time to conduct an inter-methodology comparison, focusing on logistic regression results and execution times. As shown in Table 19, the balanced accuracy of the two anonymized datasets is equivalent. Despite differences in sensitivity and specificity results, the overall prediction performance is very similar. Both methods produce data that is not identical to the original model, which had a balanced prediction accuracy of 53 percent. This outcome is intuitive, as privacy protection typically entails a trade-off with data utility, reflected in the differing prediction performance.

Moreover, considering the computational efficiency issues highlighted in various papers, the execution times of both models were measured for the chosen parameter selection. Although both execution times are fast, Table 23 demonstrates that the Mondrian algorithm took longer than the IPF algorithm. Therefore, based on this metric alone, the IPF algorithm appears more computationally efficient. Additionally, the IPF algorithm has extensive R packages, enabling a data publisher to

easily adjust parameters according to their preferences, whereas the Mondrian algorithm does not, making its implementation more complex.

Finally, both models have limitations in terms of utility due to their respective anonymization processes. The Mondrian technique results in a dataset with a different structure, while the IPF technique introduces slight changes in data distribution.

Table 23: Execution Times for Mondrian k-anonymization and IPF methods

Method	Execution_Time
Mondrian k-anonymization	1.131
IPF with Differential Privacy	0.393

Note: This table summarizes the execution times with a k=10 parameter for the Mondrian k-anonymization approach and epsilon=1 and priorn=0.1 parameters for the IPF approach.

5 Discussion

This section interprets and contextualizes the main findings of the study within the existing literature. It also offers explanations for any unexpected results and discusses the impact of study limitations while providing suggestions to overcome them in future research.

5.1 Expected findings

The primary objective of this research was to evaluate the effectiveness of k-anonymization and differential privacy approaches in preserving privacy, maintaining data utility, and managing computational complexity when applied to a large tabular dataset. The majority of the study's results conformed to theoretical expectations and were consistent with established literature for privacy preserving data publishing methods. Both methods demonstrated expected trade-offs between privacy and data utility as introduced by Dwork (2006), wherein high privacy levels corresponded to low data utility and vice versa.

In the case of Mondrian k-anonymization with relaxed partitioning, also a more pronounced trade-off was observed. The discernibility metric curve showed a sharp increase for certain k-values, indicating that the algorithm required more extensive data generalization to satisfy the k-anonymity criterion. In other words, the steep increase implies that achieving a higher level of anonymity necessitates significant sacrifices in data utility. Thus, the performance "elbow" in k-anonymity is located at lower levels of privacy protection. Practically, for data publishers this entails that if privacy is desired without disproportionate loss of usability for recipients, it is preferable to opt for a relatively low privacy protection level (i.e., $k = 5$) in this particular dataset. However, the observed trade-off plateaued at a certain point, as initial partitions and generalizations already met the k-anonymity requirement for subsequent increased k-values. This indicates that once a sufficiently high k-value is reached, additional privacy protection can be achieved without further loss in data usability. Data publishers can infer from this result that enhancing privacy protection beyond a certain threshold does not necessarily entail additional sacrifices in data utility.

Aligning with theoretical expectations, the IPF approach with differential privacy exhibited a trade-off dynamic as well. A low privacy budget effectively mitigated reidentification risk, confirming the robustness of differential privacy in preserving privacy while yielding the lowest data utility. As the privacy budget increased, this balance shifted into higher privacy budgets leading to increased reidentification risks and improved data utility. Initially, the utility increased more rapidly for lower values of ϵ. At higher ϵ values, the rate of utility improvement diminished slightly, suggesting diminishing returns in terms of utility gains. Meanwhile, the rate of decline in privacy protection did not increase as sharply. These results imply that substantial utility can be gained with small increases in the privacy budget, a characteristic beneficial for data publishers seeking to optimally balance the trade-off between privacy and utility.

Logistic regression analysis revealed discrepancies between the prediction accuracies of the original

and anonymized datasets. Structural modifications by the Mondrian k-anonymization resulted in an increased number of predictors, potentially enhancing prediction performance. The introduction of noise by the differential privacy method altered the frequency of most categorical levels while attempting to maintain statistical properties at an aggregate level, likely contributing to the changes in prediction accuracy. Furthermore, the differences in variable significance highlighted the implications of these anonymization techniques: some variables lost their significance, while others gained significance. Consequently, data miners might interpret the data differently and reach different conclusions when using anonymized data instead of the original dataset. However, the observed discrepancies in prediction performance was expected to some extent due to the added noise and generalization operations.

5.2 Unexpected findings

The analysis revealed several unexpected results. The strict partitioning approach in the Mondrian algorithm failed to achieve k-anonymity for any k-value, likely due to the high uniqueness of the data, which resulted in partitions with insufficient records to meet the k-anonymity criterion.

An additional unexpected finding was the inconsistency in the priors used for the IPF method. The ambiguous results for the varying prior parameter may stem from the fact that, beyond a certain point, the noise addition introduced to prevent zero counts might cause more harm than benefit, adversely affecting the overall data utility. Another reason could be that the prior values were too small given the high sparsity of the dataset (see Table 12). Higher prior values, with more significant ranges, might reveal a clearer trend caused by the varying parameter and prove more effective in achieving stable results for the IPF algorithm. Therefore, future research should consider selecting higher prior values and conducting a comparative analysis. Although there might be potential for improvement, my findings indicate that the low prior parameter values facilitated adequate convergence, suggesting their sufficiency in managing the high level of sparsity.

Furthermore, the IPF method demonstrated relatively high computational efficiency in terms of execution time compared to the Multidimensional Mondrian k-anonymization method, contrary to the initial hypothesis that k-anonymization would be more efficient than differential privacy implementations due to its simpler algorithmic structure (McSherry and Talwar, 2007). However, the difference may be attributed to implementation specifics and inherent differences in data structure handling by the two chosen methods. For instance, the Mondrian method involves recursive partitioning, which has a time complexity of $O(nlog(n))$ per iteration due to the need to repeatedly sort and find median values (leFevre et al., 2006). In simpler terms, this means that if the amount of data is doubled, the time required increases with a logarithmic factor rather than merely doubling. In contrast, the IPF method is based on fitting models to the margins of contingency tables and iteratively adjusting them until convergence, which tends to have a more straightforward and often faster convergence process. Additionally, the choice of relaxed partitioning allows for more flexible ways to group the data rather than strictly splitting it into two parts. This flexibility can be more

complex and take more time because it needs to carefully manage all these possible groups to maintain k-anonymity. Given these methodological choices, it is rational that the differentially private IPF model is computationally faster than the Mondrian approach.

A notable anomaly in the IPF anonymized data was the high standard error for a specific variable during logistic regression modeling. This suggests potential alterations in data distribution introduced by the anonymization process. More specifically, due to noise addition, this categorical level ended up with only two observations. The sparsity in the variable led to an inflated standard error. Thus, although the IPF model attempts to maintain the statistical properties, it is not always able to for every level. Additionally, the IPF method resulted in higher sensitivity but lower specificity compared to the Mondrian method, indicating a higher rate of false positives, which could be particularly problematic in medical diagnostics where false positives may lead to unnecessary interventions.

5.3 Limitations

Finally, this study has several limitations that must be acknowledged. The analysis was conducted on a single dataset with specific characteristics, which limits the generalizability of the findings. Subsequent studies should explore the application of these methods to diverse datasets with varying levels of sparsity and varying number of quasi-identifiers. When evaluating varying levels of sparsity, it is important to adjust the reidentification metric accordingly. This metric, which determines the number of records that are unique in both the original and synthetic datasets, should be normalized for the uniqueness in the original data. This adjustment is crucial because the number of unique records can vary significantly in sparse datasets, affecting the accuracy of reidentification risk assessment. Furthermore, datasets from other domains such as financial records, social media data, and genomic data should be examined to evaluate the robustness and versatility of the privacy-preserving methods.

Additionally, the transformation of categorical data into numeric form to apply the Mondrian k-anonymization method might limit generalizability. This transformation can introduce biases and affect the natural relationships between data attributes. Further investigations could examine the performance of k-anonymization on datasets with inherently numeric data, or explore newer techniques that incorporate user-defined generalization hierarchies, which allow for more meaningful anonymization without distorting the data's original structure.

This study also examined a limited range of parameter settings for both anonymization techniques. Consequently, it would be valuable to test a broader range of parameters to provide a more comprehensive evaluation of the methods' performance. For instance, extending the prior parameter settings, as previously discussed, could be beneficial. Also, exploring higher order contingency tables might be valuable for data recipients when performing analyses. Nonetheless, future research should carefully review the rapidly increasing computational complexity that comes with it. Additionally, exploring further levels of privacy budgets for the IPF approach with differential privacy or

experimenting with different k-values for the Mondrian approach could yield additional insights.

Moreover, this study concentrated on two specific PPDP methods: Multidimensional Mondrian k-anonymization and the IPF method with differential privacy. Specific methodological choices were made for these approaches. Future research could explore alternative techniques, such as local recoding for Mondrian k-anonymization and the use of distributions other than Laplace for noise generation in the IPF method as researchers McSherry and Talwar (2007) argued. Furthermore, while these methods are prominent in the field, they are not exhaustive. Given for example that k-anonymization is vulnerable to identity and attribute disclosure attacks, additional research could extend this work by including other privacy-preserving techniques such as l-diversity and t-closeness. L-diversity ensures that sensitive attributes have at least l "well-represented" values to prevent identity disclosure attacks. T-closeness ensures that the distribution of sensitive attributes within any equivalence class closely matches the distribution of those attributes in the overall table, thereby preventing attribute disclosure attacks. Comparing findings from these methods could provide a more comprehensive understanding of their relative strengths and weaknesses in different scenarios and might be a better opponent for the differential privacy approach.

In conclusion, while this study provides valuable insights into the application of k-anonymization and differential privacy, addressing these limitations through broader and more varied research will enhance the understanding and effectiveness of privacy-preserving data publishing methods.

6 Conclusion

The primary objective of this research was to compare k-anonymization and differential privacy and assess their efficacy in preserving privacy, maintaining data utility, and managing computational complexity when applied to a large tabular dataset containing sensitive medical records. After careful consideration, the Multidimensional Mondrian k-anonymization technique and the Iterative Proportional Fitting technique were applied to seek the answer to the question:

How do k-anonymization and differential privacy methods compare in terms of preserving privacy, maintaining data utility, and managing computational complexity when applied to sensitive healthcare datasets?

The findings of this study provide critical insights into the effectiveness and practical application of these privacy-preserving data publishing methods. First, in terms of privacy preservation, the relaxed partitioning method of Mondrian k-anonymization exhibited the anticipated trend: as privacy protection increases, utility decreases. Interestingly, this technique necessitated more aggressive data generalization for specific increments of smaller k-values to ensure each individual was indistinguishable from at least k-1 others in the dataset. Additionally, the discernability value plateaued at higher privacy levels, arguing that higher privacy can be maintained without further loss of utility. Conversely, the strict partitioning method of Mondrian k-anonymization was unable to meet the k-anonymity criterion at any level of privacy protection due to the high uniqueness of the data, which resulted in overly granular partitions with fewer than k indistinguishable records. The differentially private IPF method confirmed the trade-off theory as well and proved highly effective in mitigating reidentification risk for low privacy budgets. This underscores the robustness of differential privacy in preserving privacy.

In terms of maintaining data utility, again the expected trade-offs were observed. However, notably the IPF technique incorporating differential privacy exhibited a steep increase in utility with initial small increments in the privacy budget, followed by diminishing returns. These findings are particularly valuable for data providers in making informed decisions about the balance between privacy and utility. Specifically, when data publishers aim to release datasets with relatively higher usability while minimizing the risk of individual identification, they should consider opting for small increments in the privacy budget (i.e., $\epsilon = 0.5$ instead of $\epsilon = 0.5$). Additionally, consistent with expected outcomes of diminished data usability, logistic regression analysis revealed discrepancies in prediction accuracies between the anonymized datasets and the original data. These discrepancies were primarily due to structural data modifications from generalization in k-anonymization and adjusted frequency distributions from noise introduction by differential privacy. The differences also extended to the significance of variables, potentially compromising the data's usability, as data analysis could lead to different conclusions.

Contrary to initial expectations, the IPF method demonstrated higher computational efficiency compared to the Mondrian k-anonymization method. This efficiency, measured by the CPU time

of the anonymization process, may be attributed to differences in how each method handles data structures and optimizations specific to each approach. Specifically, the recursive partitioning of the Mondrian technique exhibits a logarithmic time complexity, and the relaxed partitioning technique necessitates evaluating more partitioning options, thereby requiring more time than the convergence process of the IPF method.

Finally, the varying priors in the IPF method led to ambiguous outcomes. For instance, with a privacy budget of 0.2, utility increased as the prior parameter increased, whereas for other privacy budgets, utility fluctuated with increasing prior. Theoretically, the prior smooths the data in cases of data sparsity by preventing zero counts, thereby improving utility. However, the results did not provide clear evidence supporting this or any other reasoning for the selected small priors. Given the high level of sparsity, higher levels of smoothing may be required. Thus, future research should investigate whether larger prior values demonstrate higher utility.

Based on these key findings, recommendations can be provided for potential stakeholders. For data publishers, such as healthcare organizations, the objective is to select the method and corresponding methodological choices based on the publishing preferences and the data characteristics. Although k-anonymization can anonymize data while maintaining data utility to a certain extent, the differentially private IPF method appears to offer the same advantages and additional benefits. First, the robust privacy guarantees provided by the differential privacy framework ward off more privacy threats than k-anonymization, especially in compliance with stringent regulations like the GDPR, making it a compelling choice for datasets requiring high levels of confidentiality. Second, the method offers a broader range of parameter options to suit the discussed publishing preferences and data characteristics, thereby providing greater flexibility. Third, the computational efficiency of the IPF approach incorporating differential privacy render it particularly suitable for large-scale data applications. Therefore, I recommend to use the IPF method with the differential privacy extension. The findings on utility gains relative to changes in privacy protection can assist in setting optimal parameters according to data publishers' preferences.

For regulators and policymakers, the study highlights the importance of promoting and potentially mandating advanced privacy-preserving techniques that can ensure data privacy without significantly compromising utility. This is particularly relevant as data-driven technologies continue to expand, necessitating robust frameworks to protect individual privacy.

For researchers and developers, the insights into the trade-offs and computational efficiencies of different methods provide a foundation for further refinement and development of the privacy-preserving data publishing techniques. As mentioned in the discussion, researchers could enhance the current IPF method by incorporating the capability to handle numeric input or by drawing noise from other distributions to better fit the data.

In conclusion, this research makes a significant contribution to the field of data privacy by providing a detailed comparative analysis of k-anonymization and differential privacy methods. The insights

gained are valuable for data publishers, policymakers, and researchers, fostering a more secure and privacy-conscious approach to data handling and dissemination in an increasingly data-driven world.

7 References

Abowd, J. M., Schneider, M. J., & Vilhuber, L. (2013). Differential privacy applications to Bayesian and linear mixed model estimation. *Journal of Privacy and Confidentiality, 5*(1).

Charest, A. S., & Hou, Y. (2016). On the meaning and limits of empirical differential privacy. *Journal of Privacy and Confidentiality, 7*(3), 53-66.

Cox, L. H. (1980). Suppression methodology and statistical disclosure control. *Journal of the American Statistical Association, 75*(370), 377-385.

Dalenius, T. (1977). Towards a methodology for statistical disclosure control. *Journal of the American Statistical Association, 72*(358), 867-874.

De Jong, M. G., Pieters, R., & Stremersch, S. (2012). Analysis of sensitive questions across cultures: An application of multigroup item randomized response theory to sexual attitudes and behavior. *Journal of Personality and Social Psychology, 103*(3), 543.

De Waal, A. G., Hundepool, A. J., & Willenborg, L. C. R. J. (1996). ARGUS: Software for statistical disclosure control of microdata. In *Proceedings of the Annual Research Conference* (p. 45). US Department of Commerce, Bureau of the Census.

Dwork, C. (2006, July). Differential privacy. In *International Colloquium on Automata, Languages, and Programming* (pp. 1-12). Berlin, Heidelberg: Springer.

Fung, B. C., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys, 42*(4), 1-53.

Garner, S. A., & Kim, J. (2019). The privacy risks of direct-to-consumer genetic testing: Case study of 23andMe and Ancestry. *Washington University Law Review, 96*(6), 1219-1266.

Art. 4 GDPR – Definitions - General Data Protection Regulation (GDPR). (2018, March 29). General Data Protection Regulation (GDPR). https://gdpr-info.eu/art-4-gdpr/

Gentry, C. (2009, May). Fully homomorphic encryption using ideal lattices. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing* (pp. 169-178).

Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence, 2*(6), 305-311.

LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2005, June). Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data* (pp. 49-60).

LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2006, April). Mondrian multidimensional k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)* (pp. 25-25). IEEE.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., & Vilhuber, L. (2008, April). Privacy: Theory meets practice on the map. In *2008 IEEE 24th International Conference on Data Engineering* (pp. 277-286). IEEE.

Majeed, A., & Lee, S. (2020). Anonymization techniques for privacy preserving data publishing: A comprehensive survey. *IEEE Access, 9*, 8512-8545.

McSherry, F., & Talwar, K. (2007). Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS)* (pp. 94-103). IEEE. https://doi.org/10.1109/FOCS.2007.66

Narayanan, A., & Shmatikov, V. (2008, May). Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)* (pp. 111-125). IEEE.

Nowok, B., Raab, G. M., & Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software, 74*(1), 1-26.

Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review, 57*, 1701-1777.

Prakash, P., Ding, J., Li, H., Errapotu, S. M., Pei, Q., & Pan, M. (2020, December). Privacy preserving facial recognition against model inversion attacks. In *GLOBECOM 2020-2020 IEEE Global Communications Conference* (pp. 1-6). IEEE.

Raab, G. M. (2022, September). Utility and disclosure risk for differentially private synthetic categorical data. In *International Conference on Privacy in Statistical Databases* (pp. 250-265). Cham: Springer.

Samarati, P. (2001). Protecting respondents identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering, 13*(6), 1010-1027.

Samarati, P., & Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.

Schneider, M. J., & Abowd, J. M. (2015). A new method for protecting interrelated time series with Bayesian prior distributions and synthetic data. *Journal of the Royal Statistical Society Series A: Statistics in Society, 178*(4), 963-975.

Schneider, M. J., Jagpal, S., Gupta, S., Li, S., & Yu, Y. (2018). A flexible method for protecting marketing data: An application to point-of-sale data. *Marketing Science, 37*(1), 153-171.

Song, S., Chaudhuri, K., & Sarwate, A. D. (2013, December). Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing* (pp. 245-248). IEEE.

Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10*(05), 571-588.

Sweeney, L. (1998). Datafly: A system for providing anonymity in medical data. *Database Security XI: Status and Prospects*, 356-381.

Wang, K., Fung, B. C., & Yu, P. S. (2007). Handicapping attacker's confidence: An alternative to k-anonymization. *Knowledge and Information Systems, 11*(3), 345-368.

What is data privacy? | IBM. (n.d.). https://www.ibm.com/topics/data-privacy.

Xiao, X., Wang, G., & Gehrke, J. (2010). Differential privacy via wavelet transforms. *IEEE Transactions on Knowledge and Data Engineering, 23*(8), 1200-1214.

Zhu, T., Li, G., Zhou, W., & Yu, S. P. (2017). Differentially private data publishing and analysis: A survey. *IEEE Transactions on Knowledge and Data Engineering, 29*(8), 1619-1638.

8 Appendix

Table 24: Summary Statistics

Variable	Level	Meaning	Frequency	Percentage
Gender	1	Male	1469	46.69
	2	Female	1677	53.31
Education	1	No formal education	10	0.32
	2	Education up to age 12	13	0.41
	3	Education up to age 14	45	1.43
	4	Education up to age 18	165	5.24
	5	Higher Education	582	18.50
	6	University	876	27.84
	7	Not disclosed	1455	46.25
SocialClass	1	Upper class	26	0.83
	2	Upper middle class	397	12.62
	3	Middle Class	1647	52.35
	4	Lower middle class	651	20.69
	5	Working class	362	11.51
	6	Lower class	63	2.00
WorkingStatus	1	Full-time job	1865	59.28
	2	Part-time (8-29h per week)	265	8.42
	3	Part-time (under 8h per week)	76	2.42
	4	Unemployed	158	5.02
	5	Sick/disabled	130	4.13
	6	Retired	393	12.49
	7	Housewife	221	7.02
	8	Student	38	1.21
MaritalStatus	1	Married	2144	68.15
	2	Single	1002	31.85
IllegalPurchase	1	Yes	1497	47.58
	2	No	1649	52.42

Note. This table summarizes statistics from the original dataset. The variables represent gender, education, social class, work status, marital status, and the illegal prescription drug purchase indicator and the adjoining column states their category levels with meaning. The frequency column shows the count of each level within a categorical variable, while the percentage column shows the proportion of occurrences of each level, expressed as a percentage of the total for that variable.

Table 25: Fisher's Exact Test Results

Variable 1	Variable 2	p-value
Gender	Education	0.0004998
Gender	SocialClass	0.0104948
Gender	WorkingStatus	0.0004998
Gender	MaritalStatus	0.0000204
Gender	IllegalPurchase	0.0001119
Education	SocialClass	0.0004998
Education	WorkingStatus	0.0004998
Education	MaritalStatus	0.6016992
Education	IllegalPurchase	0.1894053
SocialClass	WorkingStatus	0.0004998
SocialClass	MaritalStatus	0.0004998
SocialClass	IllegalPurchase	0.0009995
WorkingStatus	MaritalStatus	0.0004998
WorkingStatus	IllegalPurchase	0.0004998
MaritalStatus	IllegalPurchase	0.0463723

Note: This table illustrates the Fisher's Exact results between two variables with the corresponding p-value of the original dataset.

Table 26: Cramér's V Test Results

Variable 1	Variable 2	Cramér's V
Gender	Education	0.1031749
Gender	SocialClass	0.0687405
Gender	WorkingStatus	0.2705878
Gender	MaritalStatus	0.0764118
Gender	IllegalPurchase	0.0689054
Education	SocialClass	0.1424046
Education	WorkingStatus	0.1166968
Education	MaritalStatus	0.0383907
Education	IllegalPurchase	0.0528467
SocialClass	WorkingStatus	0.1303908
SocialClass	MaritalStatus	0.1186087
SocialClass	IllegalPurchase	0.0824298
WorkingStatus	MaritalStatus	0.1898090
WorkingStatus	IllegalPurchase	0.1043282
MaritalStatus	IllegalPurchase	0.0358007

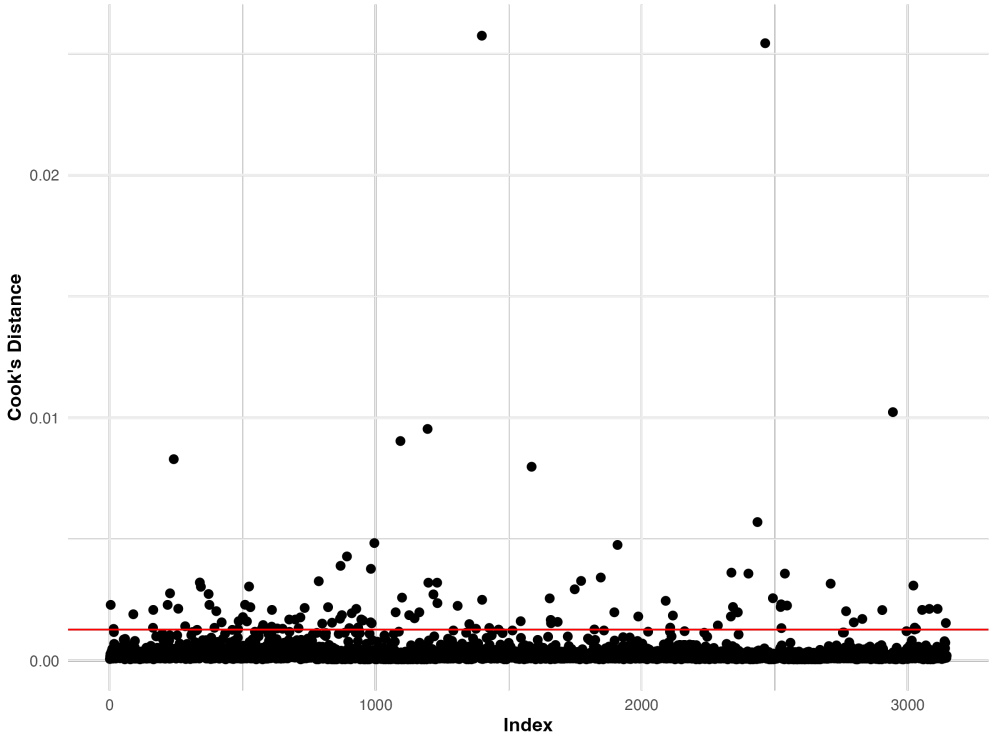
Note: This table illustrates the Cramér's V Test results between two variables with the corresponding measure of association strength in the original dataset.

Table 27: Generalized Variance Inflation Factor Results for Benchmark

	GVIF	Df	GVIF^(1/(2*Df))
Gender	1.097217	1	1.047481
Education	1.201332	6	1.015403
SocialClass	1.216420	5	1.019784
WorkingStatus	1.302417	7	1.019052
MaritalStatus	1.065341	1	1.032154

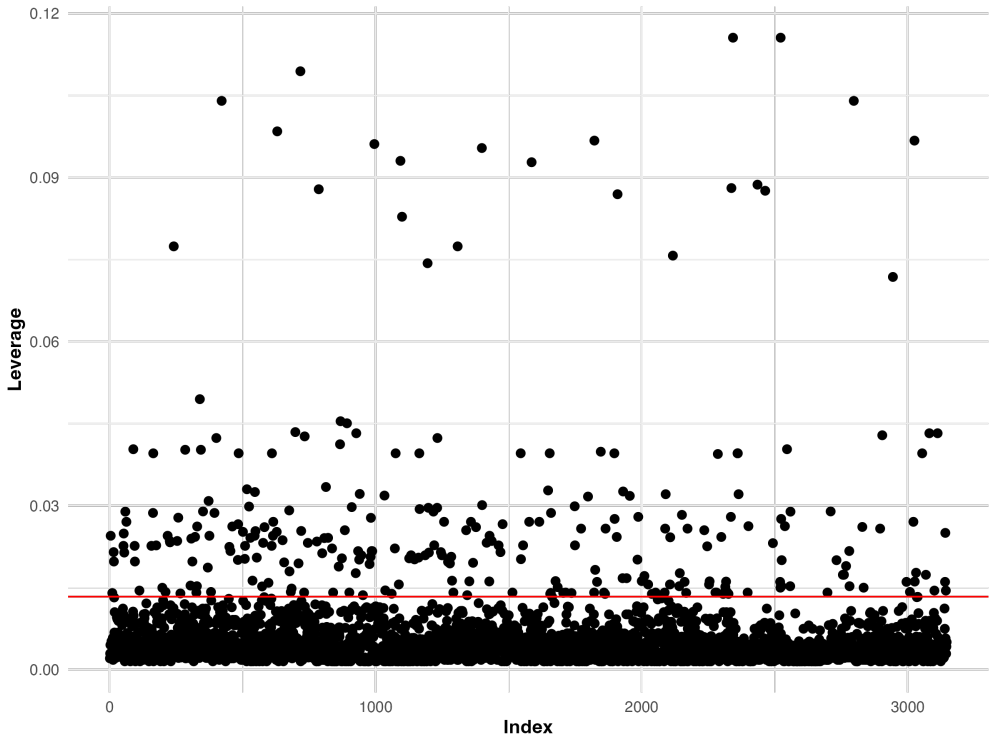
Note: This table illustrates the GVIF results for the predictor variables of the benchmark model. GVIF measures the inflation of variance of the estimated coefficients due to multicollinearity among the predictors. Due to the categorical nature of the variables, the scaled version that adjusts for the degrees of freedom associated with each predictor is also included (i.e., $\text{GVIF}^{(1/(2*DF))}$).

Figure 5: Cook's Distance



Note. The plot shows Cook's Distance for each observation in the original dataset. The horizontal red line represents the threshold for influential points, calculated as 4/(n - k - 2), where n is the number of observations and k is the number of predictors.

Figure 6: Leverage



Note. The plot shows leverage values for each observation in the original dataset. The horizontal red line represents the threshold for influential points, calculated as 2 times the mean leverage.

Table 28: Preview Mondrian Anonymized Data for k = 5

Gender	Education	SocialClass	WorkingStatus	MaritalStatus	IllegalPurchase
[1-2]	[5-7]	[2-4]	1	[1-2]	2
[1-2]	[5-7]	[2-4]	1	[1-2]	1
[1-2]	[5-7]	[2-4]	1	[1-2]	1
[1-2]	[5-7]	[2-4]	1	[1-2]	1
[1-2]	[5-7]	[2-4]	1	[1-2]	1
[1-2]	[5-7]	[2-4]	1	[1-2]	1
[1-2]	[5-7]	[2-6]	1	[1-2]	1
[1-2]	[5-7]	[2-6]	1	[1-2]	2
[1-2]	[5-7]	[2-6]	1	[1-2]	1
[1-2]	[5-7]	[2-6]	1	[1-2]	1
[1-2]	[5-7]	[2-6]	1	[1-2]	1
[1-2]	[5-7]	[2-6]	1	[1-2]	1
[1-2]	[5-7]	[2-5]	1	[1-2]	1
[1-2]	[5-7]	[2-5]	1	[1-2]	1
[1-2]	[5-7]	[2-5]	1	[1-2]	1
[1-2]	[5-7]	[2-5]	1	[1-2]	1
[1-2]	[5-7]	[2-5]	1	[1-2]	2
[1-2]	[5-7]	[2-5]	1	[1-2]	1
[1-2]	[4-7]	[2-5]	1	[1-2]	1
[1-2]	[4-7]	[2-5]	1	[1-2]	1

Note: This table provides a preview of the first 20 observations of the Multidimensional Mondrian k-anonymized data.

Table 29: Preview Mondrian Anonymized Data for k = 10

Gender	Education	SocialClass	WorkingStatus	MaritalStatus	IllegalPurchase
[1-2]	[5-7]	[2-6]	1	[1-2]	2
[1-2]	[5-7]	[2-6]	1	[1-2]	1
[1-2]	[5-7]	[2-6]	1	[1-2]	1
[1-2]	[5-7]	[2-6]	1	[1-2]	1
[1-2]	[5-7]	[2-6]	1	[1-2]	1
[1-2]	[5-7]	[2-6]	1	[1-2]	1
[1-2]	[5-7]	[2-6]	1	[1-2]	1
[1-2]	[5-7]	[2-6]	1	[1-2]	2
[1-2]	[5-7]	[2-6]	1	[1-2]	1
[1-2]	[5-7]	[2-6]	1	[1-2]	1
[1-2]	[5-7]	[2-6]	1	[1-2]	1
[1-2]	[5-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-5]	1	[1-2]	1
[1-2]	[4-7]	[2-5]	1	[1-2]	1
[1-2]	[4-7]	[2-5]	1	[1-2]	1
[1-2]	[4-7]	[2-5]	1	[1-2]	1
[1-2]	[4-7]	[2-5]	1	[1-2]	2
[1-2]	[4-7]	[2-5]	1	[1-2]	1
[1-2]	[4-7]	[2-5]	1	[1-2]	1
[1-2]	[4-7]	[2-5]	1	[1-2]	1

Note: This table provides a preview of the first 20 observations of the Multidimensional Mondrian k-anonymized data.

Table 30: Preview Mondrian Anonymized Data for k = 15

Gender	Education	SocialClass	WorkingStatus	MaritalStatus	IllegalPurchase
[1-2]	[4-7]	[2-6]	1	[1-2]	2
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	2
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	2
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1

Note: This table provides a preview of the first 20 observations of the Multidimensional Mondrian k-anonymized data.

Table 31: Preview Mondrian Anonymized Data for k = 20

Gender	Education	SocialClass	WorkingStatus	MaritalStatus	IllegalPurchase
[1-2]	[4-7]	[2-6]	1	[1-2]	2
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	2
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	2
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1
[1-2]	[4-7]	[2-6]	1	[1-2]	1

Note: This table provides a preview of the first 20 observations of the Multidimensional Mondrian k-anonymized data.

Table 32: Preview DP Synthesis

Gender	Education	SocialClass	WorkingStatus	MaritalStatus	IllegalPurchase
2	3	4	5	1	1
2	5	3	4	1	2
1	4	3	4	1	2
2	6	3	1	1	2
1	3	3	2	2	2
1	3	2	2	1	1
2	6	2	3	2	2
2	5	3	4	1	1
2	4	3	4	1	1
1	5	2	5	1	2
1	1	3	4	1	1
2	5	2	3	1	1
1	3	2	7	2	2
1	5	2	3	1	2
1	4	3	4	1	1
2	5	3	2	1	2
1	3	3	4	2	2
2	4	3	2	1	1
1	6	3	5	1	2
1	4	5	4	1	2

Note: This table provides a preview of the first 20 observations of the IPF anonymized data for epsilon=1 and priorn=0.1.

Table 33: Generalized Variance Inflation Factor Results for Mondrian

	GVIF	Df	GVIF^(1/(2*Df))
Gender	1.843040	2	1.165155
Education	2.491787	6	1.079052
SocialClass	3.295795	9	1.068503
WorkingStatus	7.925014	13	1.082871
MaritalStatus	1.626216	1	1.275232

Note: This table illustrates the GVIF results for the predictor variables of the Mondrian k-anonymized model. GVIF measures the inflation of variance of the estimated coefficients due to multicollinearity among the predictors. Due to the categorical nature of the variables, the scaled version that adjusts for the degrees of freedom associated with each predictor is also included (i.e., $\text{GVIF}^{(1/(2*DF))}$).

Table 34: Generalized Variance Inflation Factor Results for IPF

	GVIF	Df	GVIF^(1/(2*Df))
Gender	1.047093	1	1.023276
Education	1.173775	6	1.013442
SocialClass	1.088712	5	1.008536
WorkingStatus	1.142298	7	1.009548
MaritalStatus	1.022209	1	1.011043

Note: This table illustrates the GVIF results for the predictor variables for the differentially private IPF model. GVIF measures the inflation of variance of the estimated coefficients due to multicollinearity among the predictors. Due to the categorical nature of the variables, the scaled version that adjusts for the degrees of freedom associated with each predictor is also included (i.e., $\text{GVIF}^{(1/(2*DF))}$).