University of Southern Brittany

Internship Project

Evaluating Privacy Techniques for Secure Data Publishing

Author:

Laurine Owino

Supervisor

Prof. Pierre-Martin Tardiff Universite de Sherbrooke

Academic Year: 2024-2025

Preface

This report was prepared as part of my internship at the Université de Sherbrooke. Due to visa processing challenges, the internship was conducted remotely from within my countro of study, France. Despite the physical distance, regular virtual meetings were held once a week with my supervisor to ensure consistent guidance, feedback, and progress monitoring.

The motivation for this internship stemmed from my supervisor's interest in developing techniques for anonymizing data that may be shared publicly. As a result, this project was designed to serve as foundational background work by exploring and evaluating state-of-the-art Privacy Enhancing Techniques (PETs), specifically in the area of Privacy-Preserving Data Publishing (PPDP). The study focused on techniques such as anonymization, synthetic data generation, and differential privacy — all of which are essential in addressing the growing concerns around data sharing, privacy threats, and the trade-off between privacy protection and data utility.

Throughout the internship, all the planned objectives were successfully achieved, and the work plan was followed as scheduled. The remote working arrangement did not hinder the progress or the quality of the outcomes, and the experience provided valuable insight into both the theoretical and practical aspects of privacy-preserving data techniques.

Abstract

In the era of data-driven decision-making, organizations frequently collect and share large volumes of personal data, raising serious privacy concerns. Privacy-Preserving Data Publishing (PPDP) techniques aim to mitigate these risks while maintaining the utility of the shared data for analytical tasks. This study evaluates three widely adopted PPDP approaches: anonymization, synthetic data generation, and differential privacy. For anonymization, k-anonymity was applied using the AnonyPy and Anjana libraries, with privacy measured via the k parameter and utility assessed through the level of data suppression. Synthetic data generation was conducted using the SDV library (including GaussianCopula, CTGAN, and TVAE models), with privacy assessed through statistical distance metrics and utility tested via a logistic regression classifier trained on synthetic data and tested on real data. Differential privacy was implemented using IBM's Diffprivlib, where privacy was controlled through the ϵ (epsilon) parameter, and utility measured via classification accuracy.The results demonstrate a clear privacy-utility trade-off across all techniques. Differential privacy provided the best balance, preserving model performance while offering formal privacy guarantees. Synthetic data generation resulted in a notable utility loss (approximately 7% reduction in accuracy), while anonymization's utility was impacted by suppression requirements. These findings highlight that the selection of a PPDP technique depends on the specific context and the acceptable balance between privacy protection and data usability.

Keywords: PPDP, privacy, anonymization, synthetic data generation, differential privacy

Contents

List of Figures

List of Tables

Acronyms

GAN Generative Adversarial Network. IV

ML Machine Learning. IV

PET Privacy-Enhancing Techniques. IV

PPDP Privacy-Preserving Data Publishing. IV

SDG Synthetic Data Generation. IV

1 Introduction

In the modern digital economy, organizations across various sectors—including healthcare, finance, education, retail, and government—collect vast amounts of data about individuals, transactions, behaviors, and preferences. These data assets are increasingly recognized as valuable resources, driving advancements in machine learning, data-driven decision-making, policy formulation, and academic research. In many cases, there is a need to share these datasets with third parties, such as research institutions, developers, public agencies, or even the general public. Data sharing can lead to important societal benefits, such as improved healthcare solutions, fraud detection systems, and economic models, by enabling researchers and developers to analyze rich, real-world datasets.

There are several legitimate reasons why organizations may wish to share or publish data. For example, in healthcare, publicly available datasets can accelerate medical research, facilitate the development of diagnostic tools, and support epidemiological studies. In finance, shared transaction data can aid in the detection of fraudulent patterns and enhance risk assessment models. In public administration, governments may release census or employment data to promote transparency and allow independent policy evaluation. In the private sector, companies may share consumer or operational data with partners to drive innovation and improve services. Open data initiatives by public institutions also underscore the growing importance of data sharing for societal benefit.

Despite these advantages, sharing data presents significant privacy risks. Datasets may contain direct identifiers (such as names or social security numbers) as well as quasi-identifiers (such as age, gender, and ZIP code) that, when combined, can be exploited to re-identify individuals—even if obvious identifiers are removed. Several studies have shown that de-identified datasets can still be vulnerable to re-identification attacks, especially when adversaries have access to auxiliary information. The consequences of such privacy breaches can be severe, including identity theft, financial fraud, discrimination, and reputational harm.

As a result, organizations are increasingly obligated to protect personal data under various regulatory frameworks, such as the General Data Protection Regulation (GDPR) in Europe, the California Consumer Privacy Act (CCPA) in the United States, national Data Protection Acts, and other global privacy standards. These legal and ethical obligations have led to research into Privacy-Preserving Data Publishing (PPDP) techniques, which aim to enable data sharing without compromising individual privacy.

Among the most prominent and widely used PPDP approaches are anonymization, synthetic data generation, and differential privacy:

- Anonymization techniques, such as k-anonymity, l-diversity, and t-closeness, attempt to generalize, suppress, or modify dataset attributes so that individual records cannot be distinguished within a group of similar records. However, anonymization often leads to information loss, especially when strict privacy guarantees are required.
- Synthetic data generation involves creating entirely new datasets that retain the statistical properties and structure of the original data but contain no actual individual records. This approach can prevent direct linkage to real individuals, but its success depends on how well the synthetic data captures important relationships within the

original dataset.

- Differential privacy provides mathematically rigorous privacy guarantees by adding controlled noise to the data or to the outputs of learning algorithms. This ensures that the presence or absence of any single individual in the dataset does not significantly affect the analysis, making it nearly impossible to infer private information about specific data points.

The objectives of the study are:

- To explore and implement various Privacy Enhancing Techniques (PETs) used in Privacy-Preserving Data Publishing (PPDP).
- To evaluate and compare the effectiveness of these techniques in protecting individual privacy.
- To analyze the privacy-utility trade-off associated with each of the techniques in practical data publishing scenarios

In the study, the three approaches named above were evaluated using the UCI Adult Income dataset, a commonly used benchmark in machine learning and privacy-preserving research. The specific tools employed include AnonyPy and Anjana for anonymization; the SDV library with multiple synthesizers (GaussianCopula, CTGAN, and TVAE) for synthetic data generation; and IBM's Diffprivlib for applying differential privacy to a logistic regression model. These libraries and frameworks were selected due to their wide adoption and active use in both academic research and practical applications, making the findings relevant for real-world privacy-preserving data publishing scenarios.

A crucial aspect of this evaluation was to assess both the privacy protection and the data utility offered by each technique. However, since anonymization fundamentally alters the structure of data and suppresses records to achieve privacy goals, its utility could not be measured through classification tasks. Instead, the level of record suppression was used as an indicator of information loss. In contrast, the utility of synthetic data and differentially private models was evaluated using classification accuracy, where a logistic regression model trained on transformed data was tested on the original dataset. Additionally, the privacy parameter ϵ was varied in the differential privacy experiments to observe the relationship between privacy strength and model utility.

The findings of this study highlight the inherent trade-offs between privacy and utility that must be considered when choosing a PPDP technique. Notably, differential privacy achieved the best balance, maintaining strong privacy guarantees without severely degrading classification performance. Synthetic data generation resulted in an approximate 7% drop in accuracy, indicating limitations in the current state of generative models for preserving predictive patterns. Anonymization techniques required substantial suppression to satisfy privacy constraints, leading to considerable information loss and reduced dataset usability for analytical tasks.

This paper is structured as follows: Section 2 provides a review of related work on PPDP techniques; Section 3 details the methodology and experimental setup; Section 4 presents and discusses the results; and Section 5 offers concluding remarks and directions for future research.

2 Literature Review

As data becomes an increasingly valuable asset in sectors such as healthcare, finance, public administration, and technology, the need to share collected data with third parties, such as researchers, policymakers, and collaborators, has grown significantly. Public data sharing plays a vital role in fostering innovation, enabling scientific discovery, and developing new technologies such as machine learning models. However, this practice raises serious concerns regarding the protection of personal and sensitive information contained in these datasets. Even after obvious identifiers are removed, datasets may still contain subtle information that can lead to individual re-identification.

In the context of data privacy, attributes in a dataset are typically categorized as: [4] [20]

- Direct Identifiers (DI) - attributes that directly identify an individual. These can include the name, email address, social security number, etc. An example is as shown on column DI of figure 1. These are often explicitly removed during de-identification processes.
- Quasi-Identifiers - auxilliary information, that when combined can tell reveal an individual's identity. These include age, gender, ZIP code, etc, as shown on the Quasi Identifiers columns on table 1. Even though these attributes seem innocuous individually, combinations of them can significantly narrow down potential matches, leading to re-identification attacks. These are generalized or supressed.
- Sensitive Attributes - highly critical attributes, usually protected by law and regulations. An example is shown on column SA of figure 1. Normally, an individual would want this hidden. For example, religion, sexual orientation, crime disease and political opinion. These are retained as is, as they are usually the main target for analysis.
- Non-Sensitive Attributes - the other attributes that do not contribute to sensitive information. An example is height as shown on column NSA of figure 1. These are not collected, removed, published as is.

	DI	NSA	Quasi Identifiers (QIs)			SA
ID	Name	Height	Age	Zip Code	Martial Status	Crime
1	Joe	5	29	32042	Separated	Murder
2	Jill	4	20	32021	Single	Theft
3	Sue	6	24	32024	Widowed	Traffic
4	Abe	5	28	32046	Separated	Assault
5	Bob	7	25	32045	Widowed	Piracy
6	Amy	6	23	32027	Single	Indecency

Figure 1: Data Attributes [20]

Research by Latanya Sweeney famously demonstrated that 87% of the U.S. population could be uniquely identified using only ZIP code, birth date, and gender—highlighting the serious privacy risks posed by quasi-identifiers. [30]

Failure to properly handle identifiers and quasi-identifiers can expose data subjects to various privacy threats, including:

h!

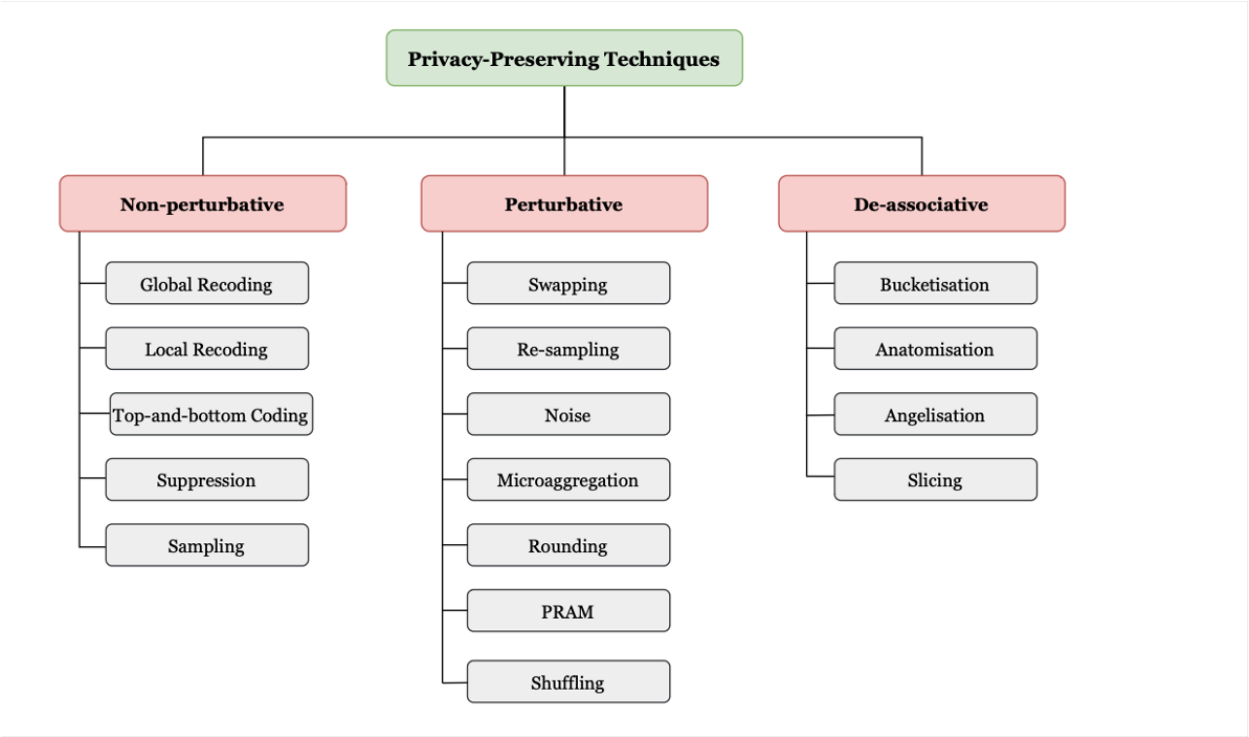

Figure 2: Privacy Preserving Techniques [4]

- Identity Disclosure: Where an adversary learns the exact identity of an individual associated with a data record.
- Attribute Disclosure: Where sensitive information about an individual (e.g., medical condition, income level) is inferred, even without explicit re-identification.
- Membership Inference: Where an adversary determines whether a particular individual is part of the dataset.

These risks underline the need for careful Privacy-Preserving Data Publishing (PPDP) techniques when releasing data. To mitigate such risks, various privacy preserving technologies have been developed. These techniques are grouped into 3 major groups, pertubative, non-pertubative and de-associative techniques. Non-pertubative techniques involve the reduction of detail or even suppression of information; pertubative techniques relates to distortion of information while de-associative techniques break the relationship between QI and sensitive attributes, releasing two separate tables instead of just one with both QI and sensitive attributes. [4]

A specific subset of privacy preserving techniques, Privacy-Preserving Data Publishing (PPDP), focuses on techniques that enable the safe release or sharing of datasets while minimizing the risk of disclosing sensitive information [11]. PPDP methods are particularly important when data needs to be publicly available or transferred to external stakeholders. Commonly used PPDP techniques include:

- Anonymization [20] - Modifying the dataset by generalizing or suppressing quasi-identifiers to prevent re-identification. Examples include k-anonymity, l-diversity, and t-closeness, which provide structured criteria for protecting against privacy attacks but can degrade data utility through information loss.

- Pseudonymization [23] - Replaces direct identifiers in a dataset with pseudonyms or artificial identifiers (such as random strings or codes) to prevent direct linkage to an individual. While this protects identity in the immediate dataset, the original information can still be retrieved if the mapping key is preserved. As a result, pseudonymization reduces re-identification risk but does not offer complete privacy if the key is compromised. It is often used as an initial step before applying stronger privacy
- Differential Privacy [8] - Providing formal privacy guarantees by introducing mathematically calibrated noise, ensuring that the inclusion or exclusion of any single individual does not significantly affect the output of computations or analyses.
- Synthetic Data Generation [25] - Creating entirely artificial datasets that mimic the statistical properties of the original data, thus preventing direct disclosure of real individual records. However, the effectiveness of synthetic data depends heavily on the quality of the generative model used.
- Homomorphic Encryption [22] - Allowing computations to be performed directly on encrypted data without the need to decrypt it first. The result, when decrypted, matches the output of the same computation performed on the raw data. This technique enables secure data processing by untrusted parties, such as cloud servers, without exposing sensitive information. While powerful, homomorphic encryption is computationally intensive and less commonly used in large-scale PPDP tasks compared to anonymization or differential privacy.

A persistent challenge in PPDP is the inherent trade-off between privacy and utility. Stronger privacy protection typically requires altering the data more aggressively—through noise addition, generalization, or suppression—leading to a loss in data utility, which can degrade the performance of downstream tasks such as machine learning or statistical analysis. Conversely, preserving data utility may leave residual privacy risks if insufficient safeguards are applied. [5]

This study explores and empirically compares these three major PPDP techniques, anonymization, synthetic data generation, and differential privacy, using the widely studied UCI Adult Income dataset. The evaluation considers both privacy protection (e.g., suppression level, distance measures, and privacy parameters like ϵ) and data utility (measured via classification performance in logistic regression tasks). The techniques are implemented using well-established libraries such as AnonyPy, Anjana, SDV, and Diffprivlib, which represent state-of-the-art tools in their respective domains.

By analyzing the strengths and weaknesses of these methods, this work aims to shed light on the practical considerations involved in choosing appropriate PPDP techniques, highlighting the real-world impact of the privacy-utility trade-off.

2.1 Anonymization

Anonymization is a fundamental approach in Privacy-Preserving Data Publishing (PPDP) that transforms data to prevent re-identification of individuals while attempting to retain its analytical value. Various models and methods guide this process to reduce privacy risks. Common methods of anonymization include [20] [11]:

- **Generalization**: Replacing specific data values with broader categories (e.g., replacing exact ages with age ranges). An example in a datset would be changing the age of 22 to <25 or 20-25.
- **Suppression**: Omitting certain attribute values or entire records to prevent identification risks. The value is usually replaced by '*'. For example, the age 22 would be 2*.
- **Perturbation**: Modifying data by adding noise or slightly altering attribute values to mask the original information. Sometimes this involves replacing values with synthetically generated values
- **Permutation**: Records partitioned into groups and shuffled within those groups.
- **Anatomization**: Involves separating the quasi-identifiers from the sensitive attributes thus breaking the association between the 2. The data is then released as 2 different tables

These methods can be used individually or in combination to satisfy formal privacy models like k-anonymity, l-diversity, and t-closeness.

2.1.1 k-anonymity

The k-anonymity model ensures that each record in the dataset is indistinguishable from at least k-1 other records with respect to a set of quasi-identifiers. Thus, the probability of re-identifying someone becomes 1/k. While this model reduces identity disclosure risk, it does not guard against attribute inference attacks. The table shown in figure 3 shows a dataset that is k-anonymous.
The method works by placing at least k users in an equivalence class (EC) with same QI's values.It is commonly achieved using generalization and suppression, balancing privacy with data utility-[31] [20].

	Quasi Identifiers		Sensitive Attribute
ID	Age	Country	Political views
1	35	Greenland	Liberal
2	35	Canada	Conservative
3	38	Belize	Liberal
4	40	Belize	Liberal
5	37	Canada	Conservative
6	37	Canada	Conservative

(a) Original data about the users (six records).

	Quasi Identifiers		Sensitive Attribute
ECs	Age	Country	Political views
C_1	35-37	North America	Liberal
	35-37	North America	Conservative
C_2	38-40	Central America	Liberal
	38-40	Central America	Liberal
C_3	35-37	North America	Conservative
	35-37	North America	Conservative

(b) 2-anonymous data about the users (i.e., $k = 2$).

Figure 3: K-Anonymity [20]

2.1.2 l-diversity

To enhance protection against inference attacks, l-diversity requires that each equivalence class contains at least l distinct, well-represented values for the sensitive attribute, preventing adversaries from confidently guessing sensitive information. It is hard to achieve l-diversity without over-generalizing data in sparse datasets and inference attacks can still occur in semantically similar sensitive attribute values such as different chronic illnesses. [19]. An example of dataset which conforms to l-diversity is shown in figure 4 l-diversity is of 3 types:

- Distinct - Each group must have at least l distinct values for the sensitive attribute.
- Entropy - Uses the concept of entropy (information theory) to ensure the distribution is not too skewed, preventing a single value from dominating even if there are technically l distinct values.
- Recursive - Prevents cases where one sensitive value is much more frequent than the others, making it possible for an attacker to make high-confidence guesses.

	Quasi Identifiers		Sensitive Attribute
ID	Age	Country	Political views
1	35	Greenland	Liberal
2	35	Canada	Conservative
3	37	Canada	Conservative
4	37	Canada	Conservative

(a) Original data about the users (QIs and SA).

	Quasi Identifiers		Sensitive Attribute
ECs	Age	Country	Political views
C_1	35-37	North America	Liberal
	35-37	North America	Conservative
	35-37	North America	Conservative
	35-37	North America	Conservative

(b) 2-diverse data about the users (i.e., $l = 2$).

Figure 4: l-diversity [20]

2.1.3 t-closeness

The t-closeness model addresses limitations of l-diversity by ensuring that the distribution of sensitive attributes in any group closely resembles the overall data distribution (within a threshold t). The model measures the distance between an SA in an EC and the global distribution and ensures it is less than or equal to a threshold t. [17] An example of data that is t-close anonymous is shown in figure 5. This reduces the risk of inferential attacks based on distributional differences.

2.1.4 Summary

Table 1 shows a summary of the anonymization techniques discussed in this section.

	Quasi Identifiers		Sensitive Attribute
ID	Zip code	Age	Disease
1	47677	29	Gastric ulcer
2	47602	22	Gastritis
3	47678	27	Stomach cancer
4	47905	43	Gastritis
5	47909	52	Flu
6	47906	47	Bronchitis
7	47605	30	Bronchitis
8	47673	36	Pneumonia
9	47607	32	Stomach cancer

(a) Original data about the users (QIs and SA).

		Quasi Identifiers		Sensitive Attribute
ECs	ID	Zip code	Age	Disease
C_1	1	4767*	<40	Gastric ulcer
	3	4767*	<40	Stomach cancer
	8	4767*	<40	Pneumonia
C_2	4	4790*	>40	Gastritis
	5	4790*	>40	Flu
	6	4790*	>40	Bronchitis
C_3	2	4760*	<40	Gastritis
	7	4760*	<40	Bronchitis
	9	4760*	<40	Stomach cancer

(b) t-close anonymous data (i.e., $t = 0.278$).

Figure 5: t-closeness [20]

Model	Protects Against	Measures	Weaknesses
k-anonymity [31]	Identity disclosure	Frequency of quasi-ID values	Vulnerable to attribute disclosure
ℓ-diversity [19]	Attribute disclosure	Count/entropy of sensitive values	Can't handle semantic similarity or skewed data
t-closeness [17]	Stronger attribute disclosure	Distance between distributions	Harder to implement, can reduce data utility

Table 1: Comparison of Privacy Models in Data Anonymization

2.2 Synthetic Data Generation

Synthetic data generation is an emerging Privacy-Preserving Data Publishing (PPDP) technique that involves producing entirely artificial data that mimics the statistical properties and relationships of the original dataset. This approach allows organizations to share data without exposing actual sensitive information, thus mitigating risks of re-identification and disclosure [25]. Unlike anonymization—which modifies or masks real data—synthetic data generation creates new records from scratch based on models learned from the original data. This process ensures that no real individual's information is present in the synthetic dataset, making it resistant to identity or attribute disclosure attacks. Some common techniques for generating synthetic data include [25]:

- GAN-based [10] - Originally based on two different artificial neural networks trained simultaneously in a competitive manner. Currenntly, noise is added to the models to ensure privacy.

- Machine Learning-based [7]- Uses a structure based on machine learning to generate synthetic data. Models may be based on Neural Networks (NN), Deep Learning (DL) or linear regression.
- Statistical based - Uses a structure based on statistics which learns the underlying patterns and distributions of the data and generates new data samples that meet specific needs or conditions. Examples include marginal, stochastic, histogram, and density-based algorithms.
- Kernel based - Uses a structure based on a kernel function, such as a Gaussian, on each data point and summing them up to obtain a smooth curve that approximates the true density function.

While synthetic data offers strong privacy protection by design, maintaining data utility—the ability for the synthetic data to support meaningful analysis or machine learning tasks—is a key challenge [29]. If the synthetic data diverges too far from the original distribution, its usefulness may be compromised. Conversely, overly accurate synthetic data may inadvertently leak patterns from the real data, increasing privacy risks. Generally, utility is considered more than relevance. [25]. Despite the strong privacy, synthetically generated data is still susceptible to membership inferential attacks. To beef up its privacy, differentially privacy is being used in synthetically generated data.

In terms of metrics, there is no consensus on how to measure the level of privacy of synthetically generated data, though measures of distance, such as Euclidean distance, have been suggested as adequate means [14]. Performance of such models are measured through statistics, utility, distance similarity. Utility is typically measured by applying machine learning models (e.g., classification) trained on synthetic data and testing their performance on real data [18].

2.3 Differential Privacy

Differential Privacy (DP) is a mathematical framework designed to provide strong privacy guarantees when analyzing and sharing data. Unlike anonymization or synthetic data generation, differential privacy focuses on limiting the risk of disclosing information about any single individual, regardless of an attacker's background knowledge. It ensures that the output of a computation does not reveal too much about any single individual's data, even if attackers have access to auxiliary information by introducing noise to a random query. [32] At its core, differential privacy ensures that the output of a computation (such as a query result or a machine learning model) does not significantly change when any single individual's data is added or removed from the dataset. This is typically formalized using two parameters [8]:

- ϵ (epsilon): Represents the privacy loss budget; smaller values indicate stronger privacy.
- δ (delta): Allows for a small probability of the privacy guarantee being violated.

Differential privacy is achieved through these techniques:

- Output Perturbation: Adding noise to the result of queries, such as counts, sums, or averages.
- Input Perturbation: Adding noise directly to the dataset features before analysis.

- Model Perturbation: Incorporating noise during machine learning model training (e.g., noisy gradients).

By adding carefully calibrated random noise to computations or training processes, DP masks the contribution of individual data points, protecting them from inference attacks.

Differential privacy is of 2 main types:

- Local [24] [6] - The data curator is not trusted thus each user perturbs their own data before sharing, useful in distributed systems. Examples of where such models are used include Google Chrome's RAPPOR and Apple's usage statistics sharing.
- Global [8] [9] - Assumes a trusted curator adds noise before releasing results. An example of where this was used was the US 2020 Census [1]

A fundamental challenge in applying differential privacy is balancing privacy and utility. Higher privacy (lower ϵ) requires more noise, reducing data utility. Lower privacy (higher ϵ) allows more accurate results but weaker privacy protection. This trade-off must be carefully managed based on the intended application and sensitivity of the data. [2]

2.4 Utility vs. Privacy Trade-offs

A fundamental challenge in Privacy-Preserving Data Publishing (PPDP) is balancing privacy protection with data utility. While stronger privacy guarantees are essential to prevent risks such as identity or attribute disclosure, they often come at the cost of reducing the dataset's usefulness for analysis, modeling, or decision-making tasks. [5]

In anonymization techniques like k-Anonymity, achieving higher privacy (by increasing the value of k) typically requires more generalization or suppression of data, which reduces the granularity and specificity of information available for downstream tasks [3]. This can lead to poorer model performance, reduced analytical insights, or data inconsistency.

In synthetic data generation, the process of creating new data points that mimic the statistical distribution of the original dataset inherently introduces approximation errors. While privacy is enhanced because individual records from the real dataset are not disclosed, utility may decline if the synthetic data fails to accurately preserve complex relationships or patterns critical for predictive tasks, as evidenced by reduced model accuracy in some cases. [29]

With differential privacy, the injection of carefully calibrated noise into data queries or model training processes guarantees formal privacy protection (controlled via the parameter). However, higher privacy (smaller values) generally results in noisier outputs, potentially degrading the accuracy or interpretability of models trained on such data. [2]

Thus, the trade-off between privacy and utility is inherent and technique-dependent. The appropriate balance depends on the data's intended use: scenarios requiring strict privacy (e.g., sensitive health data) may tolerate reduced utility, while applications demanding high data accuracy (e.g., financial forecasting) may accept weaker privacy guarantees.

2.5 Summary of Literature Review

In this study, various Privacy Enhancing Techniques (PETs) applicable to Privacy-Preserving Data Publishing (PPDP) were reviewed, each addressing the challenge of protecting sensitive information while maintaining data utility. Classical anonymization approaches such as k-Anonymity, l-Diversity, and t-Closeness were discussed, which aim to prevent identity and

attribute disclosure through generalization, suppression, and perturbation. However, these methods can lead to significant information loss and are vulnerable to background knowledge attacks.

Synthetic data generation was also examined as a promising alternative that creates entirely new datasets mirroring the statistical properties of real data. Different methods based on different techniques including statistics, machine-learning and GANs were highlighted, offering varying degrees of success in maintaining data utility and privacy.

Finally, Differential Privacy (DP) was explored as a mathematically rigorous framework providing quantifiable privacy guarantees by introducing calibrated noise during data analysis or model training. Despite its strong privacy protection, DP methods often involve a trade-off between privacy (controlled by the ϵ parameter) and utility.

Overall, the literature reveals that each technique involves inherent trade-offs between privacy and data utility, and their effectiveness depends on the specific context, data characteristics, and the intended use of the published data.

3 Methodology

The experimental evaluation in this study focuses on three key privacy-preserving data publishing (PPDP) techniques: anonymization, synthetic data generation, and differential privacy. Each technique was implemented using an open-source Python library, selected based on prior academic research that demonstrated its effectiveness in handling privacy-preserving tasks. The libraries chosen for this work — AnonyPy, Anjana, SDV, and Diffprivlib — have been validated and applied in recent academic studies focusing on privacy-preserving methods for tabular data processing [27] [28] [21] [14] [12]. These studies informed the selection to ensure that the tools used align with best practices and yield comparable and reliable results. All experiments and evaluations were implemented in Python for this study (see Appendix A for source code details).

3.1 Dataset

The experiments in this study were carried out using the UCI Adult Income dataset from Kaggle [16], a widely used benchmark dataset for evaluating machine learning models and privacy-preserving techniques. The dataset used comprises 32,561 instances and 14 attributes, including continuous and categorical variables such as age, class of work, education, marital status, occupation, and hours per week. The target variable represents whether an individual's income exceeds $50,000 per year as a binary classification: <=50K or >50K. An overview of the data is as shown in Figure 6

Before applying privacy-preserving techniques, the dataset was subjected to pre-processing steps to ensure consistency and compatibility with the applied methods. Categorical features were label-encoded, and records with missing values were replaced with the mode(categorical features) to prevent inaccuracies in privacy transformation processes. The resulting cleaned dataset was used in all subsequent experiments.

	age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country	income
0	90	?	77053	HS-grad	9	Widowed	?	Not-in-family	White	Female	0	4356	40	United-States	<=50K
1	82	Private	132870	HS-grad	9	Widowed	Exec-managerial	Not-in-family	White	Female	0	4356	18	United-States	<=50K
2	66	?	186061	Some-college	10	Widowed	?	Unmarried	Black	Female	0	4356	40	United-States	<=50K
3	54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct	Unmarried	White	Female	0	3900	40	United-States	<=50K
4	41	Private	264663	Some-college	10	Separated	Prof-specialty	Own-child	White	Female	0	3900	40	United-States	<=50K

Figure 6: Dataset Overview

3.2 Tools and Libraries Used

3.2.1 Anonypy

AnonyPy [13] is an anonymization is an open-source Python library available on Github that uses the Mondrian algorithm. The algorithm partitions data into smaller groups, each containing at least k records — this ensures that no individual record is distinguishable from at least k-1 others. Its effectiveness has been demonstrated in anonymization tasks for structured data [27] [28]. The library allows the following privacy preserving techniques.

- k-anonymity
- l-diversity
- t-closeness

3.2.2 Anjana

Anjana [28] is a multi-model library allows the application of different anonymity techniques based on a set of identifiers, quasi-identifiers (QI) and a sensitive attribute. It's built on pyCANON and uses hierarchical generalization + optional suppression tailored to each privacy model. his tool has been applied in recent research on data privacy solutions [14]. Privacy models that can be applied using this library are:

- k-anonymity
- (α ,k)-anonymity.
- l-diversity.
- Entropy l-diversity.
- Recursive (c,l)-diversity.
- t-closeness.
- Basic β -likeness.
- Enhanced β -likeness.
- δ-disclosure privacy.

3.2.3 Synthetic Data Vault (SDV)

The SDV library [26], built by Patki et al, works by using these 3 steps:

- Extracts all relevant information from the dataset and transforming the contents into numerical values using the *DataNavigator*.
- Creating generative models using the *Modeler*
- Generating synthetic data using the *Sampler*

The library uses different models to generate synthetic data. The most common include the Gaussian Copula, which is statistical based and the fastest; TVAE Synthesizer is based on neural networks; and CTGAN is neural-network based.

3.2.4 Diffprivlib

Diffprivlib [15] is an open-source and general-purpose library by IBM used to explore the effect of differential privacy on machine learning and data analytics. Diffprivlib's implementation has been validated in previous studies focusing on privacy-preserving machine learning [12]. Diffprivlib is made up of four major components:

- Mechanisms - the building blocks of differential privacy, and are used in all models that implement differential privacy.
- Models - the machine learning models with differential privacy. These include clustering, classification, regression, dimensionality reduction and pre-processing, similar to scikit-learn models.

- Tools - generic tools for differentially private data analysis such as histograms, which follow the same format as Numpy's histogram function
- Accountant - used to track privacy budget and calculate total privacy loss using advanced composition techniques.

3.2.5 scikit-learn

Scikit-learn is an open-source library built on NumPy, SciPy, and matplotlib widely used for predictive data analysis. It has various algorithms for classification, regression and clustering including logistic regression, linear regression, support-vector machines, random forests, k-means and DBSCAN. In addition, it features data pre-processing methods and utility methods such as splitting data for training and testing.

3.3 Anonymization

Anonymization was performed using both AnonyPy and Anjana libraries. The goal was to achieve k-Anonymity on selected quasi-identifiers: age, education, marital status, occupation, sex and native country. Experiments were conducted for k values of 2, 3, 5, 10, 50 and 100, representing increasing levels of privacy.
AnonyPy applied generalization and suppression to the quasi-identifiers to meet the k-Anonymity criteria.
Anjana implemented a microaggregation approach that grouped records into clusters and replaced attribute values with aggregated representations to reduce the risk of re-identification.

3.4 Synthetic Data Generation

Synthetic data was generated using the SDV library, which provides multiple synthesizer models for tabular data. During preliminary testing, the following models were evaluated:

- GaussianCopula (selected for final experiments)
- CTGAN (Conditional Tabular GAN)
- TVAE (Tabular Variational AutoEncoder)

The GaussianCopula synthesizer was ultimately chosen due to its superior balance of data utility and simplicity. A synthetic dataset of the same size as the original (32,561 records) was produced. The quality of data was measured using SDV's inbuilt tool. To assess privacy, the statistical distance(Euclidean distance) between the original and synthetic data distributions was measured. Data utility was evaluated by training a logistic regression classifier on the synthetic data and testing it on the original data to assess generalization.

3.5 Differential Privacy

Differential privacy was incorporated using Diffprivlib, applying differentially private logistic regression models to the dataset. The experiments tested the effect of varying the privacy budget (ϵ) with the values $\epsilon = 0.1, 0.2, 0.3, 0.5, 1.0, 5.0, 10.0$.

These settings allowed exploration of the trade-off between privacy strength and data utility. Model training parameters followed scikit-learn defaults, except where altered by diffprivlib's privacy constraints.

3.6 Utility Evaluation

The utility of the transformed datasets was evaluated by performing a binary classification task (predicting income level) using Logistic Regression from scikit-learn.

- Train/Test Split: The dataset was divided into a 70/30 stratified split.
- Metrics Used: Classification Accuracy and F1-Score were measured to assess predictive performance.
- For synthetic data, the model was trained on synthetic records and tested on real data to examine generalization.
- For differential privacy, both training and testing were performed on real or transformed data as appropriate.
- For anonymization, the prediction task was not performed as the data was transformed to a different version. Instead, level of supression was used as a gauge for utility.

Privacy of synthetic data was assessed by calculating statistical distances between original and synthetic feature distributions, providing a measure of dissimilarity as a proxy for disclosure risk

3.7 Summary of Experimental Setup

Table 2 shows a summary of the experimental setup.

Table 2: Summary of Techniques, Libraries, Parameters, and Evaluation Metrics

Technique	**Libraries Used**	**Parameters**	**Privacy Measurement**	**Utility Measurement**
Anonymization	AnonyPy, Anjana	k = 2, 3, 5, 10, 50, 100	k value (anonymity parameter)	Suppression level (percentage of records suppressed)
Synthetic Data Generation	SDV	GaussianCopula as main model; CTGAN and TVAE tested	Distance between real and synthetic data	Classification accuracy (train on synthetic, test on real data)
Differential Privacy	Diffprivlib	$\epsilon = 0.1, 0.5, 1, 5, 10$	Epsilon (ϵ) parameter	Classification accuracy (on real test data)

4 Results and Discussion

This section discusses the results obtained from the experiments on the dataset.

4.1 Evaluation Metrics

To assess the impact of the applied privacy-preserving techniques, two key aspects were evaluated, data utility and privacy assessment.

Data utility was measured using classification performance on the binary income prediction task. Logistic Regression classifiers were trained on transformed data and evaluated on real data (or appropriate subsets). The main performance metric used was Accuracy, supplemented by F1-score to account for class imbalance.

Privacy Assessment was carried out as follows:

- For synthetic data, privacy was estimated by computing the statistical distance (Euclidean distance) between the distributions of original and synthetic datasets.
- For differential privacy, privacy strength was controlled through the privacy budget parameter ϵ, with lower values offering stronger privacy.
- For anonymization, privacy level was controlled via k-values (3, 5, 10).

4.2 Anonymization Results

Anonymization carried out using Anonypy produced multi-dimensional columns which were considered unusable for this study(see Appendix B. Using Anjana, data utility vs privacy was measured through increasing the value of k (thus increasing privacy) and noting the level of suppression of records. It was noted that higher values of k resulted in higher values of supression as in Table 3.

k	Suppressed Records (%)	Suppressed QIs
2	30.75	0
3	42.53	0
4	27.31	0
5	31.26	0
10	43.71	1
50	37.78	3
100	38.42	4

Table 3: Suppression statistics for different values of k

4.3 Synthetic Data Generation Results

Synthetic datasets generated by GaussianCopula, CTGAN, and TVAE models were evaluated. The GaussianCopula synthesizer was selected due to its simplicity, modifiability and time taken for generation as shown in Table 4.

Synthesizer	Quality	Time Taken (s)
GaussianCopula	0.8442	4.3
CTGAN	0.8553	744.3
TVAE	0.8995	156.6

Table 4: Comparison of synthesizers in terms of data quality and time taken.

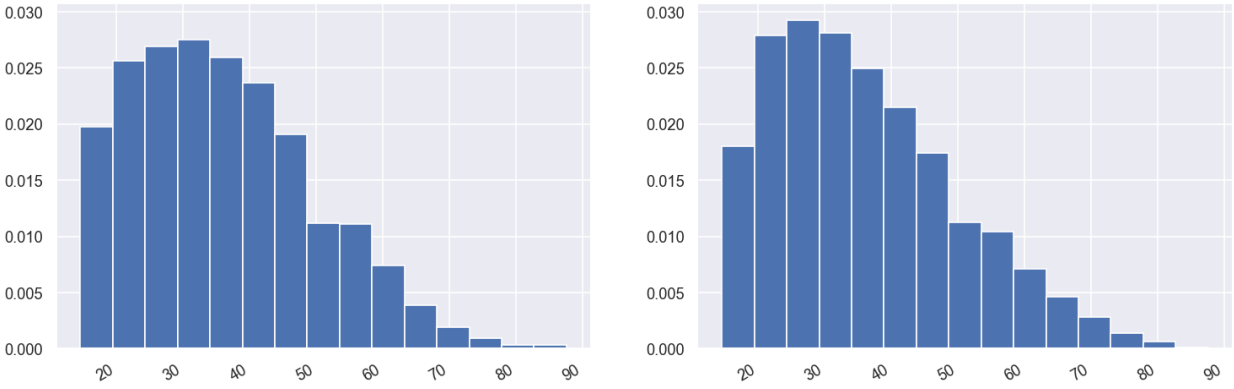

Figure 7: Histogram of Age in Original vs Synthetic Data

To further evaluate quality, histograms of the original and synthetic data were analyzed to confirm whether the data follows the statistical trends in the original data. Examples for age (Figure 7), workclass (Figure 8) and capital gain (Figure~9) are shown below. The full histogram differences are found in the Appendix C.

A correlation matrix to compare mutual information was also done which shows that the data is similar but not the same as shown in Figure 10 Data utility was measured through the performance of logistic regression on real vs synthetic data. It is observed that there is a drop of about 7% in accuracy on the synthetic data. Table 5 shows the classification report for the model trained and tested in real data.

Privacy analysis was done through measurement of Euclidean distance between the different rows. This ensures that records in the synthetic data are not the same as on original data, which would lead to privacy concerns. The results shown in Figure 11 show that most values are between 1 and 3; which implies that the data is not the same.

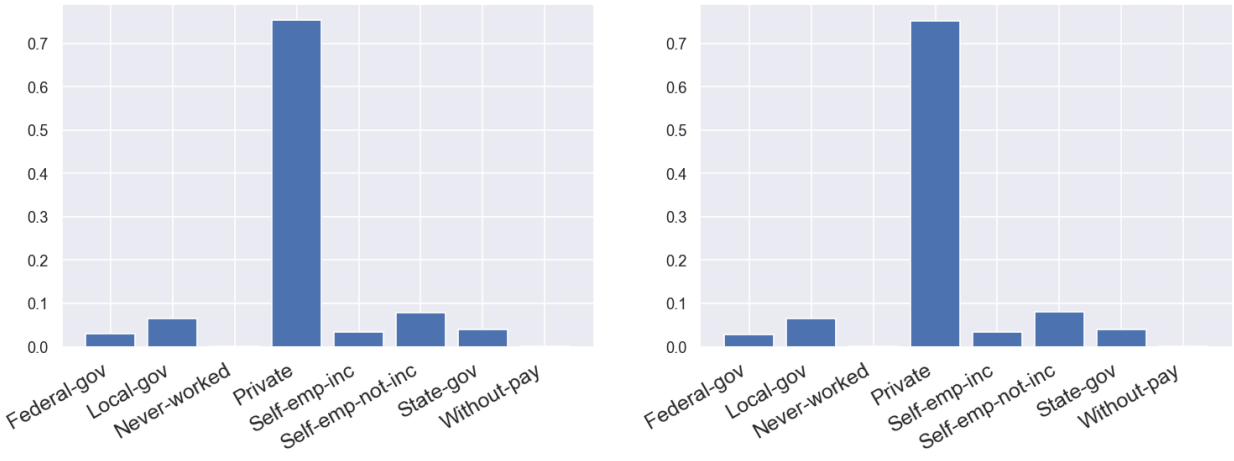

Figure 8: Histogram of Workclass in Original and Synthetic Data

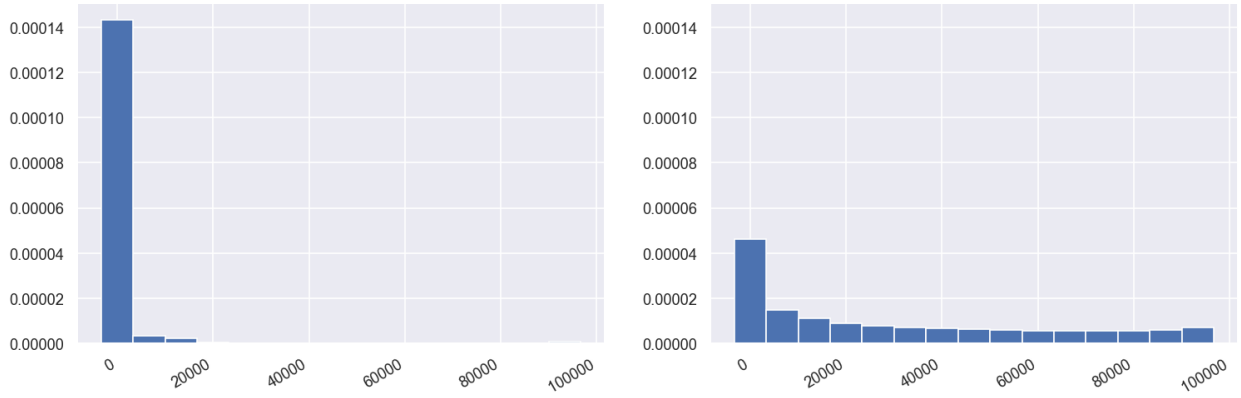

Figure 9: Histogram of Capital Gain in Original vs Synthetic Data

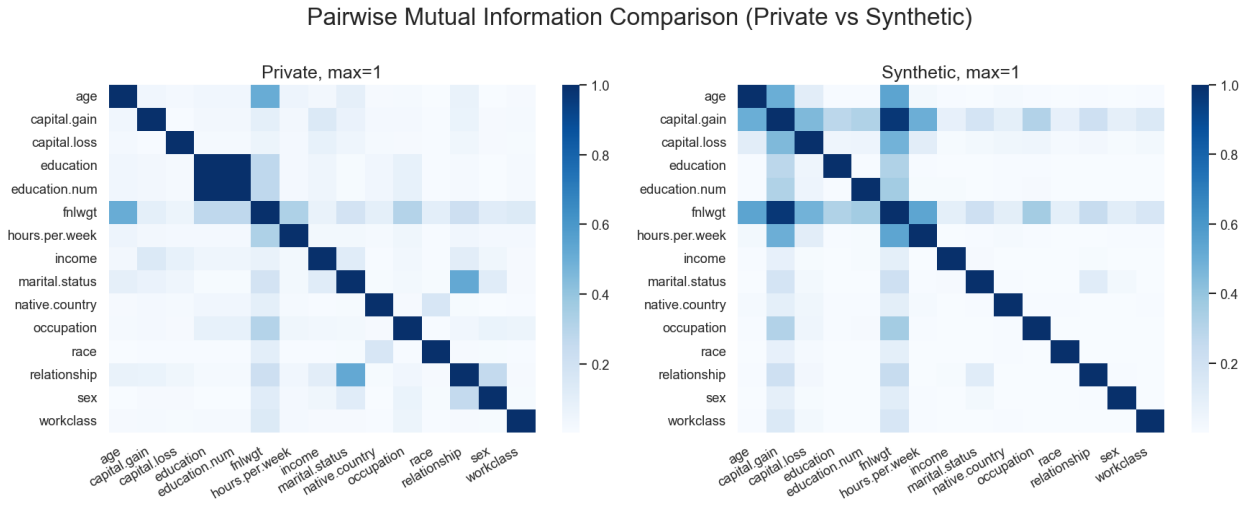

Figure 10: Correlation Matrix Between Original and Synthetic Data

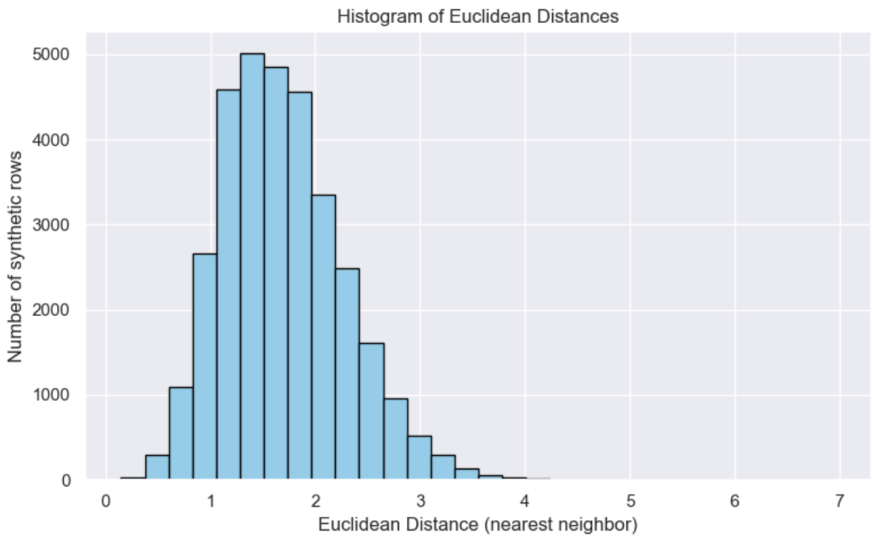

Figure 11: Histogram of Euclidean Distances between Real and Synthetic Data

Class	Precision	Recall	F1-Score	Support
$\leq 50K$	0.85	0.94	0.89	7429
$> 50K$	0.72	0.45	0.55	2340
Accuracy			0.83	9769
Macro Avg	0.78	0.70	0.72	9769
Weighted Avg	0.81	0.83	0.81	9769

Table 5: Classification report for the model trained and tested on real data.

Class	Precision	Recall	F1-Score	Support
$\leq 50K$	0.77	1.00	0.87	7429
$> 50K$	0.90	0.08	0.14	2340
Accuracy			0.78	9769
Macro Avg	0.83	0.54	0.51	9769
Weighted Avg	0.80	0.78	0.70	9769

Table 6: Classification report for the model trained on synthetic data.

4.4 Differential Privacy Results

The impact of differential privacy on model performance at varying privacy budgets (ϵ) is presented in Table **??**.

Epsilon (ϵ)	Classification Accuracy
0.1	0.7652
0.2	0.8145
0.3	0.8121
0.5	0.8239
1.0	0.8272
5.0	0.8270
10.0	0.8265

Table 7: Classification accuracy for different values of ϵ in the Differential Privacy model.

A strong privacy guarantee ($\epsilon = 0.1$) caused a significant drop in model performance (about 6%), while a higher ϵ value (weaker privacy) allowed for accuracy closer to that of the original dataset (0.8259). It is observed that further increase in ϵ results in a very small change in accuracy. This illustrates the well-known privacy-utility trade-off inherent in differential privacy mechanisms.

4.5 Discussion

Anonymization methods such as those provided by AnonyPy and Anjana demonstrated that suppression or generalisation of data reduces usability of the data. For example, it proved difficult to perform classification tasks on the data compared to the original data because the shape of the data changed. Anjana outperformed AnonPy in preserving utility due to its selective generalization strategy. It is observed that AnonyPy creates a dataset with mutliple values in each column, which may be unusable depending on the downstream task. However, Anjana requires the definition of hierarchies beforehand. It can be hypothe-

sized that an increase in the value of k increases the level of supression and thus reduces utility.

Synthetic data generation via GaussianCopula resulted in the fastest and most customizable data generation among the synthetic models tested. Although CTGAN and TVAE offer better data quality, the methods are much slower and generally harder to debug, thus not considered for this use case. The data generated is considered usable and follows most of the statistical trends in the original dataset without being the same thus protecting privacy.

Differential Privacy (using Diffprivlib) showed a clear privacy-utility trade-off controlled by ϵ. Small ϵ values degraded utility substantially, while higher ϵ values produced models that approximated the baseline performance without privacy constraints. However, a further increase beyond 5.0 seems to yield no better improvement to the performance of the classification task.

In general, differential privacy synthetic data generation with GaussianCopula provided the best balance of privacy and utility in this study, while differential privacy offered strong formal guarantees at the cost of reduced model accuracy under strict privacy settings.

5 Conclusion

This project investigated Privacy Enhancing Techniques (PETs) and compared three prominent privacy-enhancing techniques—anonymization, synthetic data generation, and differential privacy—in the context of secure data publishing using the UCI Adult Income dataset. These techniques were selected because of their broad usage in both academic research and industry practice.

To assess the impact of these techniques, two key aspects were examined: privacy guarantees and data utility. However, since anonymization fundamentally transforms the dataset structure, direct classification-based utility assessment was not possible for this method. Instead, the suppression level of records—the proportion of data entries removed or masked to satisfy privacy constraints—was used as a proxy measure of utility loss. Higher suppression levels indicated greater information loss and lower utility.

For synthetic data generation using the SDV library (GaussianCopula, CTGAN, and TVAE synthesizers), classification-based utility evaluation revealed an approximate 7% reduction in model accuracy compared to real data. This suggests that while synthetic data can protect privacy by breaking direct links to real records, capturing complex data patterns sufficiently for downstream tasks remains challenging with current generative models.

In contrast, differential privacy, implemented using IBM's Diffprivlib for logistic regression, demonstrated the best overall trade-off between privacy and utility in this study. With appropriately tuned privacy budgets (ϵ), the method preserved much of the model's predictive performance while offering formal and quantifiable privacy guarantees. However, stronger privacy guarantees came at the cost of reduced model accuracy.

Overall, the study demonstrated that no single technique is universally superior; the best method depends on the specific privacy requirements and acceptable utility loss for the application. While each technique presented clear advantages, they also highlighted distinct limitations. Anonymization led to information loss through suppression; synthetic data reduced utility in classification tasks; and differential privacy required careful parameter tuning to avoid excessive noise injection.

A limitation of this study is the reliance on a single dataset and a single machine learning model (logistic regression) for evaluation. Additionally, the privacy evaluation of synthetic data remains an open challenge, as commonly used statistical distances may not fully reflect disclosure risk.

For future work, expanding the analysis to multiple datasets and model types could yield more generalized insights. Exploring other methods of measuring utility across different techniques would also provide better insights. In addition, researching methods of measuring privacy as well as modelling an attacker would provide better understanding on the effect of PETs on data.

References

[1] John M Abowd. The us census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2867–2867, 2018.

[2] Alberto Blanco-Justicia, David Sánchez, Josep Domingo-Ferrer, and Krishnamurty Muralidhar. A critical review on the use (and misuse) of differential privacy in machine learning. *ACM Computing Surveys*, 55(8):1–16, 2022.

[3] Justin Brickell and Vitaly Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 70–78, 2008.

[4] Tânia Carvalho, Nuno Moniz, Pedro Faria, and Luís Antunes. Survey on privacy-preserving techniques for data publishing. *arXiv preprint arXiv:2201.08120*, 2022.

[5] Prathamesh P Churi and Ambika V Pawar. A systematic review on privacy preserving data publishing techniques. *Journal of Engineering Science & Technology Review*, 12(6), 2019.

[6] Graham Cormode, Somesh Jha, Tejas Kulkarni, Ninghui Li, Divesh Srivastava, and Tianhao Wang. Privacy at scale: Local differential privacy in practice. In *Proceedings of the 2018 international conference on management of data*, pages 1655–1658, 2018.

[7] Ashish Dandekar, Remmy AM Zen, and Stéphane Bressan. A comparative study of synthetic dataset generation techniques. In *Database and Expert Systems Applications: 29th International Conference, DEXA 2018, Regensburg, Germany, September 3–6, 2018, Proceedings, Part II 29*, pages 387–395. Springer, 2018.

[8] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

[9] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

[10] Alvaro Figueira and Bruno Vaz. Survey on synthetic data generation, evaluation methods and gans. *Mathematics*, 10(15):2733, 2022.

[11] Benjamin Fung, ke Wang, Rui Chen, and Philip Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 42, 06 2010.

[12] Gonzalo Munilla Garrido, Joseph Near, Aitsam Muhammad, Warren He, Roman Matzutt, and Florian Matthes. Do i get the privacy i need? benchmarking utility in differential privacy libraries. *arXiv preprint arXiv:2109.10789*, 2021.

[13] glassonion1. Anonypy. `https://github.com/glassonion1/anonypy`, 2023. Accessed: May 13, 2025.

[14] Markus Hittmeir, Andreas Ekelhart, and Rudolf Mayer. On the utility of synthetic data: An empirical evaluation on machine learning tasks. In *Proceedings of the 14th international conference on availability, reliability and security*, pages 1–6, 2019.

[15] Naoise Holohan, Stefano Braghin, Pól Mac Aonghusa, and Killian Levacher. Diffprivlib: the ibm differential privacy library. *arXiv preprint arXiv:1907.02444*, 2019.

[16] UCI Machine Learning. Adult census income. `hhttps://www.kaggle.com/datasets/uciml/adult-census-income`, 2016. Accessed: May 15, 2025.

[17] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd international conference on data engineering*, pages 106–115. IEEE, 2006.

[18] Yingzhou Lu, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, and Wenqi Wei. Machine learning for synthetic data generation: a review. *arXiv preprint arXiv:2302.04062*, 2023.

[19] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *Acm transactions on knowledge discovery from data (tkdd)*, 1(1):3–es, 2007.

[20] Abdul Majeed and Sungchang Lee. Anonymization techniques for privacy preserving data publishing: A comprehensive survey. *IEEE access*, 9:8512–8545, 2020.

[21] Mattia Mazzoli, Janis Elfert, Paolo Sacerdoti, Marco Hirsch, Michael Davis Tira, and Daniela Paolotti. Assessing synthetic data generation utility for cohort data secondary use. *medRxiv*, pages 2025–06, 2025.

[22] Manish M Potey Mr, Chandrashekhar A Dhote, and Deepak H Sharma Mr. Homomorphic encryption for security of cloud data. *Procedia Computer Science*, 79:175–181, 2016.

[23] Thomas Neubauer and Johannes Heurix. A methodology for the pseudonymization of medical data. *International journal of medical informatics*, 80(3):190–204, 2011.

[24] Thông T Nguyên, Xiaokui Xiao, Yin Yang, Siu Cheung Hui, Hyejin Shin, and Junbum Shin. Collecting and analyzing data from smart device users with local differential privacy. *arXiv preprint arXiv:1606.05053*, 2016.

[25] Pablo A Osorio-Marulanda, Gorka Epelde, Mikel Hernandez, Imanol Isasa, Nicolas Moreno Reyes, and Andoni Beristain Iraola. Privacy mechanisms and evaluation metrics for synthetic data generation: A systematic review. *IEEE Access*, 2024.

[26] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *2016 IEEE international conference on data science and advanced analytics (DSAA)*, pages 399–410. IEEE, 2016.

[27] Muhammad Ariiq Ramadhan and Nur Aini Rakhmawati. Safeguarding student data privacy: A comparative study of anonymization techniques using the mondrian algorithm. In *2024 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, pages 13–18. IEEE, 2024.

[28] Judith Sáinz-Pardo Díaz and Álvaro López García. An open source python library for anonymizing sensitive data. *Scientific data*, 11(1):1289, 2024.

[29] Fatima Jahan Sarmin, Atiquer Rahman Sarkar, Yang Wang, and Noman Mohammed. Synthetic data: Revisiting the privacy-utility trade-off. *International Journal of Information Security*, 24(4):1–22, 2025.

[30] Latanya Sweeney. Uniqueness of simple demographics in the us population. *LIDAP-WP4, 2000*, 2000.

[31] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570, 2002.

[32] Ying Zhao, Jia Tina Du, and Jinjun Chen. Scenario-based adaptations of differential privacy: A technical survey. *ACM Computing Surveys*, 56(8):1–39, 2024.

Appendices

A Source Code

The complete source code developed for this project is available at the following GitHub repository:

`https://github.com/lam0y/privacy-techniques-ppdp.git`

B Anonymization

This section shows figures related to anonymization of data using k-anonymity.

	age	education	marital.status	occupation	native.country	income	count
0	[18-36]	[Doctorate,10th,12th,Prof-school,Preschool,7th...	[Never-married,Divorced]	[Sales,Prof-specialty,Other-service,Handlers-c...	[Nicaragua,Taiwan,Honduras,Puerto-Rico,Hong,Ja...	<=50K	18
1	[18-36]	[Doctorate,10th,12th,Prof-school,Preschool,7th...	[Never-married,Divorced]	[Sales,Prof-specialty,Other-service,Handlers-c...	[Nicaragua,Taiwan,Honduras,Puerto-Rico,Hong,Ja...	>50K	1
2	[22-36]	[Assoc-voc,1st-4th,Masters,Assoc-acdm]	[Never-married,Divorced]	[?,Adm-clerical,Priv-house-serv,Transport-moving]	[Poland,France,Jamaica,Guatemala,Haiti,Japan,I...	<=50K	11
3	[21-35]	[Assoc-acdm,1st-4th,11th,Assoc-voc]	[Married-civ-spouse,Married-spouse-absent]	[Adm-clerical,Other-service,Exec-managerial,Ma...	[Honduras,England,Thailand,Guatemala,Haiti,Vie...	<=50K	10
4	[21-35]	[Assoc-acdm,1st-4th,11th,Assoc-voc]	[Married-civ-spouse,Married-spouse-absent]	[Adm-clerical,Other-service,Exec-managerial,Ma...	[Honduras,England,Thailand,Guatemala,Haiti,Vie...	>50K	1

Figure 12: Data Overview after Anonymization with Anonypy (k=10)

	index	age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country	income
0	5	[30, 35]	Private	216864	HS-grad	9	Divorced	Other-service	Unmarried	*	Female	0	3770	45	United-States	<=50K
1	20	[35, 40]	Private	188774	Bachelors	13	Never-married	Exec-managerial	Not-in-family	*	Male	0	2824	40	United-States	>50K
2	23	[50, 55]	Private	153870	Some-college	10	Married-civ-spouse	Transport-moving	Husband	*	Male	0	2603	40	United-States	<=50K
3	24	[60, 65]	?	135285	HS-grad	9	Married-civ-spouse	?	Husband	*	Male	0	2603	32	United-States	<=50K
4	33	[50, 55]	Private	123011	Bachelors	13	Divorced	Exec-managerial	Not-in-family	*	Male	0	2559	50	United-States	>50K

Figure 13: Data Overview after Anonymization with Anjana (k=10

C Synthetic Data Generation

This section shows figures and graphs related to synthetic data generation.

	age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country	income
0	50	Self-emp-not-inc	71334	HS-grad	12	Never-married	Prof-specialty	Husband	White	Male	41262	124	33	United-States	>50K
1	36	Private	56011	Some-college	7	Married-civ-spouse	Sales	Unmarried	White	Male	75748	452	33	United-States	<=50K
2	32	Private	105622	10th	12	Never-married	Sales	Own-child	White	Female	27	3	38	United-States	<=50K
3	42	Private	253428	HS-grad	11	Never-married	Craft-repair	Husband	White	Male	6830	0	37	United-States	<=50K
4	37	Private	254280	1st-4th	13	Divorced	Prof-specialty	Husband	White	Female	61249	0	37	United-States	<=50K
5	29	State-gov	347689	HS-grad	10	Married-civ-spouse	Craft-repair	Husband	White	Male	1688	0	53	United-States	<=50K
6	60	Private	77693	Bachelors	8	Never-married	Prof-specialty	Husband	White	Female	26354	94	27	United-States	<=50K
7	46	Private	92018	Some-college	8	Married-civ-spouse	Craft-repair	Own-child	White	Female	75268	0	13	United-States	<=50K

Figure 14: Synthetic Data Overview

C.1 Histograms

The following section shows standardized histograms of the different columns in the dataset used for this study. The histograms show the distribution of the original versus the synthetic dataset generated by SDV library.

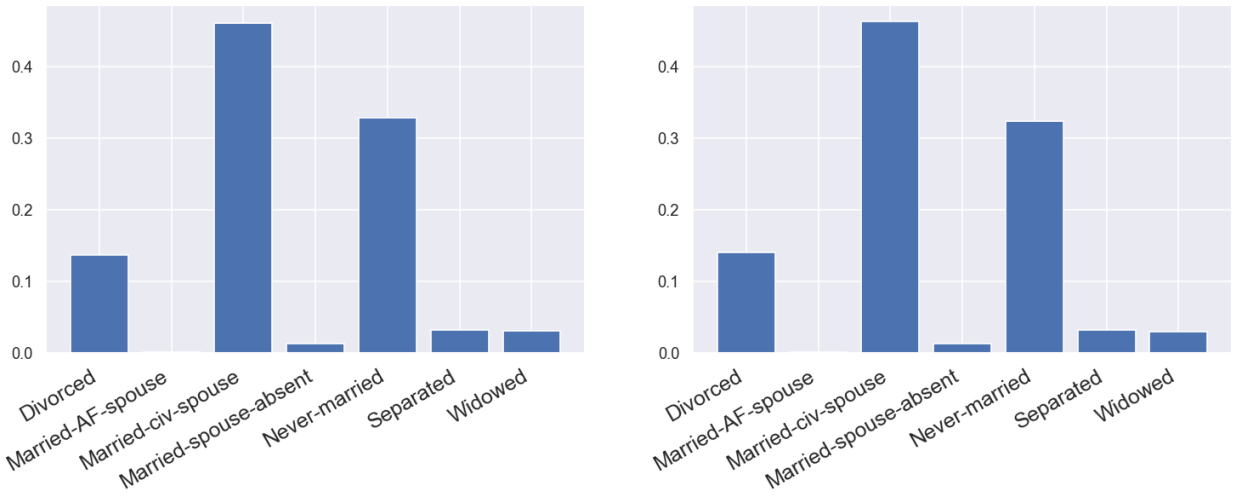

Figure 15: Marital Status

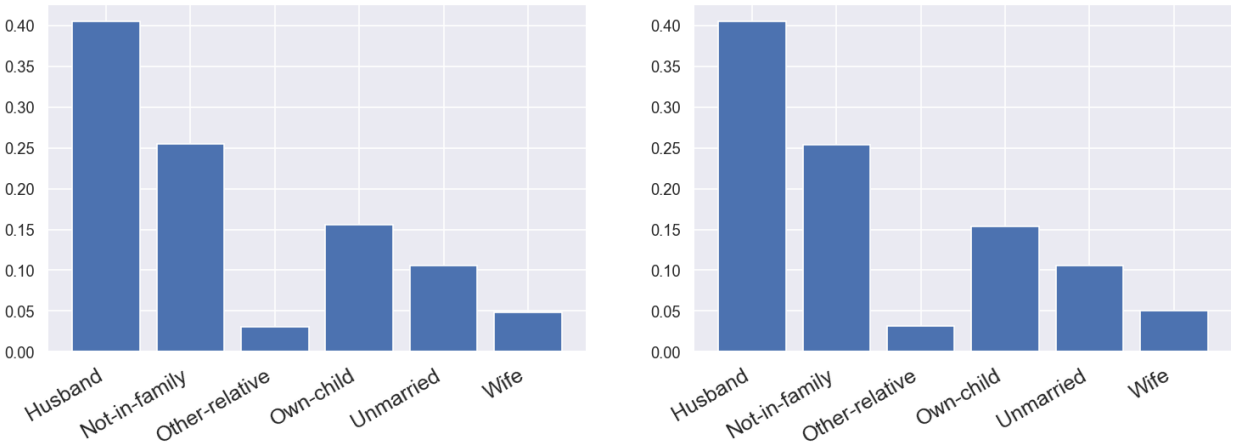

Figure 16: Relationship Status

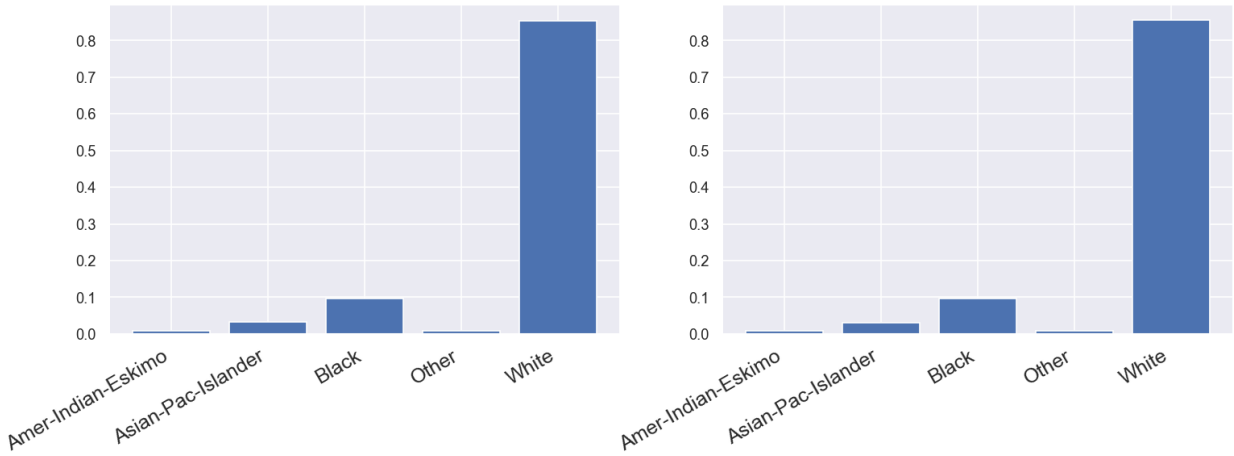

Figure 17: Race

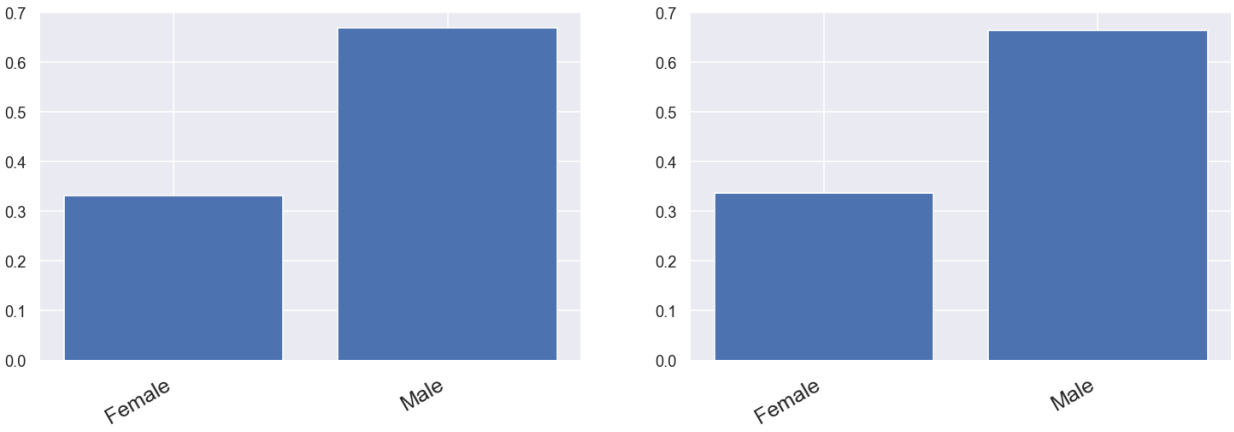

Figure 18: Gender

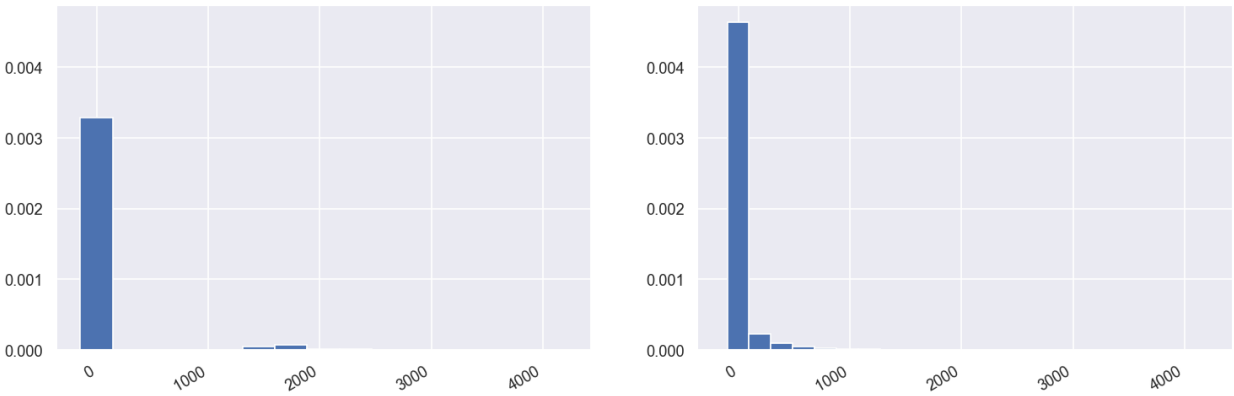

Figure 19: Capital Loss

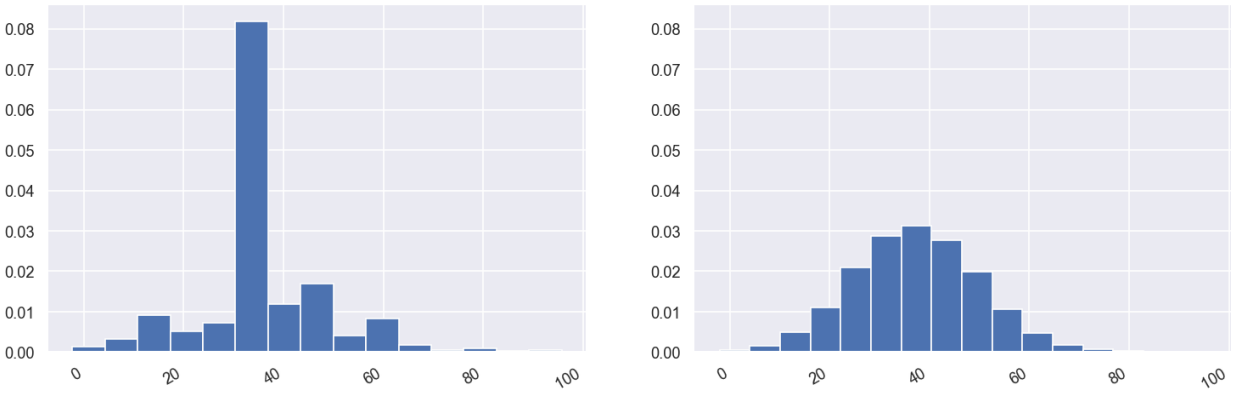

Figure 20: Hours per Week