

United Arab Emirates University

Scholarworks@UAEU

Theses

Electronic Theses and Dissertations

6-2024

VISUALIZING PRIVATELY PROTECTED DATA: EXPLORING THE PRIVACY-UTILITY TRADE-OFFS

Sarah Hayi Alkaabi

Follow this and additional works at: https://scholarworks.uaeu.ac.ae/all_theses



Part of the [Information Security Commons](#)

MASTER THESIS NO. 2024: 61**College of Information Technology****Department of Information Systems and Security****VISUALIZING PRIVATELY PROTECTED DATA:
EXPLORING THE PRIVACY-UTILITY TRADE-OFFS*****Sarah Hayi Alkaabi******June 2024***

United Arab Emirates University
College of Information Technology
Department of Information Systems and Security

**VISUALIZING PRIVATELY PROTECTED DATA: EXPLORING
THE PRIVACY-UTILITY TRADE-OFFS**

Sarah Hayi Alkaabi

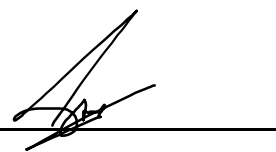
This thesis is submitted in partial fulfilment of the requirements for the degree of Master
of Science in Information Technology Management

June 2024

Cover: Visualizing Privacy: A Graphical Representation of Data Privacy Measures.
(Photo: <https://i.postimg.cc/2yLHWc5p/A-professional-and-modern-background-for-a-Power-Po.jpg>)

Declaration of Original Work

I, Sarah Hayi Alkaabi, the undersigned, a graduate student at the United Arab Emirates University (UAEU), and the author of this thesis entitled “*Visualizing Privately Protected Data: Exploring the Privacy-Utility Trade-offs*”, hereby, solemnly declare that this is the original research work done by me under the supervision of Dr. Mohamed Sharaf, in the College of Information Technology at UAEU. This work has not previously formed the basis for the award of any academic degree, diploma or a similar title at this or any other university. Any materials borrowed from other sources (whether published or unpublished) and relied upon or included in my thesis have been properly cited and acknowledged in accordance with appropriate academic conventions. I further declare that there is no potential conflict of interest with respect to the research, data collection, authorship, presentation and/or publication of this thesis.

Student's Signature: 

Date: 13/June/2024

Approval of the Master Thesis

This Master Thesis is approved by the following Examining Committee Members:

- 1) Advisor (Committee Chair): Mohamed Sharaf

Title: Associate Professor

Department of Computer Science and Software Engineering

College of Information Technology


Signature  Date 17/06/2024

- 2) Member: Rafat Damseh

Title: Assistant Professor

Department of Computer Science and Software Engineering

College of Information Technology

Signature  Date 21/06/2024

- 3) Member (External Examiner): Mohamed Adel Serhani

Title: Professor

Department of Information Systems, College of Computing and Informatics

Institution: Sharjah University, Sharjah, United Arab Emirates

Signature  Date 18/06/2024

This Master Thesis is accepted by:

Acting Dean of the College of Information Technology: Dr. Fekri Kharbash

Signature  _____

Date 05/08/2024

Dean of the College of Graduate Studies: Professor Ali Al-Marzouqi

Signature  _____

Date August 07, 2024

Abstract

In a data-driven era, achieving a balance between privacy and utility is crucial. Organizations often utilize data for research, analysis, and enhancement of services, which emphasizes the significance of effective privacy-preserving techniques to protect individuals' privacy and comply with regulations. This equilibrium is vital in data visualization to derive insightful decisions from data representations. The goal is to evaluate the trade-off between privacy preservation and data utility, understanding how differentially private parameters impact effective visualizations. Valuable insights will guide strategies for achieving optimal privacy-preserving visualization techniques. The study aims to investigate the effects on privacy and data utility in different privacy settings and identify the ideal trade-off between privacy protection and data usability. Differential privacy techniques are used to obfuscate sensitive data, relying on data sensitivity and varied epsilon values to achieve different privacy levels. The impact on data utility is analyzed using histograms, which show data frequency and distribution. These graphic aids visualized the compromises that must be made between protecting privacy and guaranteeing data usability. Our analysis highlights the vital role of visual analytics in balancing privacy protection and data utility. We present a hybrid multi-objective metric in our study that comprehensively assesses the trade-offs between privacy and utility. This novel contribution helps to develop strategies that maximize data utility and privacy protection by offering a more nuanced understanding of how privacy-preserving techniques affect data visualization. Additionally, we design an evaluation model using both empirical and estimated experiments that provide practical privacy parameters adapted for real-world scenarios. This dual-method approach offers useful techniques for protecting privacy in data visualization and determining the best trade-off, which fills a void in the existing literature. This contribution to the field delivers valuable insights to best optimize privacy-preserving techniques for data visualization.

Keywords: Data privacy, data visualization, differential privacy, privacy parameters, data utility, visual analytics, hybrid multi-objective metric, privacy-utility trade-off, histograms, evaluation model.

Title and Abstract (in Arabic)

تصور البيانات المحمية بشكل خاص: استكشاف تضارب الخصوصية والفائدة

الملخص

في عصر مدفوع بالبيانات، يعد تحقيق التوازن بين الخصوصية والفائدة أمرًا حاسمًا. غالبًا ما تستخدم المؤسسات البيانات للبحث والتحليل وتحسين الخدمات، مما يؤكد أهمية تقنيات الحفاظ على الخصوصية الفعالة لحماية خصوصية الأفراد والامتثال للوائح. هذا التوازن ضروري في تصور البيانات لاستخلاص قرارات مستنيرة من تمثيلات البيانات. الهدف هو تقييم الموازنة بين الحفاظ على الخصوصية وفائدة البيانات، وفهم كيفية تأثير المعلومات الخاصة بالخصوصية التفاضلية على التصورات الفعالة. ستوفر الأفكار القيمة استراتيجيات لتحقيق تقنيات تصور البيانات المثلى مع الحفاظ على الخصوصية. تهدف الدراسة إلى التحقيق في تأثيرات إعدادات الخصوصية المختلفة على الخصوصية وفائدة البيانات وتحديد التوازن المثالي بين حماية الخصوصية واستخدامية البيانات. تُستخدم تقنيات الخصوصية التفاضلية لتعتيم البيانات الحساسة، اعتمادًا على حساسية البيانات وقيم إيسيلون المختلفة لتحقيق مستويات خصوصية متنوعة. يتم تحليل تأثير تقنيات الحفاظ على الخصوصية على فائدة البيانات من خلال استخدام الرسوم البيانية الشريطية. بشكل خاص، تعتبر الرسوم البيانية الشريطية أدوات حيوية في التحليلات البصرية لأنها تساعد في فهم تكرار وتوزيع البيانات. هذه الأدوات الرسومية توضح التوازنات التي يجب القيام بها بين حماية الخصوصية وضمان استخدامية البيانات. تسلط تحليلاتنا الضوء على الدور الحيوي للتحليلات البصرية في موازنة حماية الخصوصية وفائدة البيانات. نقدم في دراستنا مقياسًا هجينًا متعدد الأهداف يوفر تقييمًا شاملاً للتوازنات بين الخصوصية والفائدة. تسهم هذه الإضافة الجديدة في تطوير استراتيجيات تزيد من فائدة البيانات وحماية الخصوصية من خلال تقديم فهم أكثر تفصيلاً لكيفية تأثير تقنيات الحفاظ على الخصوصية على تصور البيانات. بالإضافة إلى ذلك، نصمم نموذج تقييم باستخدام تجارب يدوية وآلية توفر معلومات خصوصية عملية مكيفة للسيناريوهات الواقعية. يقدم هذا النهج الثنائي طرقًا فعالة لحماية الخصوصية في تصور البيانات وتحديد أفضل توازن، مما يملأ فجوة في الأدبيات الحالية. تسهم هذه الإضافة في المجال بتقديم رؤية قيمة لتحسين تقنيات الحفاظ على الخصوصية في تصور البيانات.

مفاهيم البحث الرئيسية: خصوصية البيانات، تصور البيانات، الخصوصية التفاضلية، معلومات الخصوصية، فائدة البيانات، التحليلات البصرية، مقياس هجين متعدد الأهداف، توازن الخصوصية والفائدة، الرسوم البيانية الشريطية، نموذج التقييم.

Acknowledgements

I would like to thank my committee for their guidance, support, and assistance throughout my preparation of this thesis, especially my advisor Dr. Mohamed Sharaf, for his invaluable feedback and encouragement.

Special thanks go to my family and friends for their unwavering support throughout this journey.

Thank you all for being a part of this important milestone in my academic career.

Dedication

To my beloved parents and family

Table of Contents

Title	i
Declaration of Original Work	iii
Approval of the Master Thesis.....	iv
Abstract.....	vi
Title and Abstract (in Arabic)	vii
Acknowledgements.....	viii
Dedication	ix
Table of Contents	x
List of Tables	xii
List of Figures	xiii
List of Abbreviations	xiv
Chapter 1: Introduction	1
1.1 Overview	1
1.2 Problem Statement.....	3
1.2.1 Research Questions	4
1.3 Research Objectives and Contributions.....	4
Chapter 2: Related Work	6
2.1 Foundations and Practical Applications of Differential Privacy	6
2.2 Visualization Challenges in Privacy-Preserving Data Analysis.....	9
2.3 Evaluation Metrics and Methodologies	10
2.4 Navigating the Conflict: Privacy Needs vs. Utility	13
Chapter 3: Methods.....	16
3.1 Overview	16
3.2 Applying Data Privacy	17
3.3 Privacy Metrics.....	20
3.4 Generating Visualizations/Histograms	25
3.5 Accuracy/SSE Metrics.....	27
3.6 Hybrid Metric	30
3.7 Optimizing Privacy Setting	32
Chapter 4: Results and Discussions	34

4.1 Data Collection	34
4.1.1 Description of the Dataset.....	34
4.1.2 Data Preprocessing.....	35
4.2 Overview of the Main Findings.....	36
4.3 Applying Laplace Noise for Differential Privacy	37
4.4 Visualization	39
4.5 Evaluation Metrics.....	49
4.5.1 Applying Privacy Gain.....	50
4.5.2 Applying NMI.....	52
4.5.3 Empirical Exploration Results	55
Chapter 5: Conclusion	75
5.1 Managerial Implications.....	76
5.2 Research Implications	77
References.....	78

List of Tables

Table 1: Differentially Private Ages - Before and After Noise	8
Table 2: Summary Table of Existing Work and Their Comparison	15
Table 3: Perturbed Ages of Individuals at Various Epsilon Values	38
Table 4: Privacy Gain Results for education.num	52
Table 5: NMI Results for Different Epsilon Values	53
Table 6: Best Epsilon Values for Each Attribute	60
Table 7: Optimal Epsilon Values for Different Privacy and Utility Weights Across Attributes	66
Table 8: Best Epsilon Values for PG and SSE in Estimated Experiment.....	71
Table 9: Overall Best Epsilon Values for PG and SS in Estimated Experience.....	72

List of Figures

Figure 1: Effect of Differential Privacy on Evaluation Metrics.	2
Figure 2: Overview of the Research.	16
Figure 3: Handling Out-of-Range Noise Values in Laplace Perturbation.....	20
Figure 4: Effect of Epsilon Values on Age Perturbation: A Box Plot Analysis	40
Figure 5: Effect of Epsilon Values on Education Numbers Perturbation: A Box Plot Analysis	41
Figure 6: Comparison of Perturbed Education Numbers at Different Epsilon Values	42
Figure 7: Comparison of Perturbed hours per work at Different Epsilon.....	44
Figure 8: Comparison of Perturbed age at 0.5 Epsilon Values.....	45
Figure 9: Comparison of Perturbed hours per week at 0.5 Epsilon Values.....	46
Figure 10: Comparison of Perturbed capital gain at different Epsilon Value	47
Figure 11: Comparison of Perturbed capital loss at different Epsilon Value	48
Figure 12: Privacy Gain by Epsilon for Different Variables	50
Figure 13: NMI by Epsilon for Different Variables	54
Figure 14: Hybrid Metric for PG and SSE for hours.per.work.....	56
Figure 15: Hybrid Metric for PG and SSE for hours.per.work in Different Combinations of Weights	58
Figure 16: Optimal Epsilon Value Determination for education.num Using Combined Metric.....	62
Figure 17: Optimal Epsilon Value Determination for education.num Using Combined Metric for Different Combinations of Weights	64
Figure 18: Combined Metric vs Epsilon for Different Privacy and Utility Weights in Estimated Experiment (Attribute: Education Number).....	69

List of Abbreviations

DP	Differential privacy
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability Act
NMI	Normalized Mutual Information
SSE	Sum Squared Error

Chapter 1: Introduction

1.1 Overview

The study explores the complexity of striking a balance between privacy preservation and data utility in the visualization frameworks. An in-depth exploration of the main concepts of privacy preservation, data utility, and differential privacy must be conducted to ensure a comprehensive understanding. The principal issue discussed is bringing together the double consideration of the importance of prioritizing privacy as well as efficient data visualization. With the use of a case study, the research offers an insightful example of these ideas, showing the feasibility of the suggested techniques.

This study is significant because it can help close the gap between data utility and privacy preservation, which is a problem that is becoming more and more relevant in today's data-driven society. This research addresses the crucial need to preserve individual privacy while preserving the analytical value of datasets by concentrating on differential privacy. Maintaining this equilibrium is crucial for facilitating reliable and efficient data-driven decision-making in a variety of fields, such as the social sciences, finance, and healthcare.

There will be a variety of approaches employed to address the problem of striking a balance between data utility and privacy protection in visualization. Maintaining privacy is crucial to safeguarding private data and facilitating insightful data analysis. Differential Privacy is a strategy that adds randomness to the data to make it difficult to identify individual data points. The epsilon parameter, which quantifies the privacy guarantee, is a crucial element of differential privacy. It regulates the trade-off between privacy and data utility, with a larger epsilon providing better data utility at the expense of privacy and a smaller epsilon providing stronger privacy but less accuracy. Laplace noise perturbation is one of these methods that is particularly well-known and can provide practical solutions by varying the epsilon parameters. By carefully calibrating noise from a Laplace distribution and adding it to each value in the dataset, this method distorts important information while maintaining overall utility (Zhou et al., 2023). The level of privacy protection can be tailored the level of privacy protection to different requirements. In the

context of data visualization, histograms are essential tools for comprehending the frequency and distribution of data. Laplace noise is applied to histograms to illustrate the trade-offs between preserving data utility—that is, keeping the data accurate and useful—and guaranteeing privacy protection. This method emphasizes how crucial it is to identify effective privacy-preserving data visualization strategies to strike the ideal compromise between privacy and usefulness.

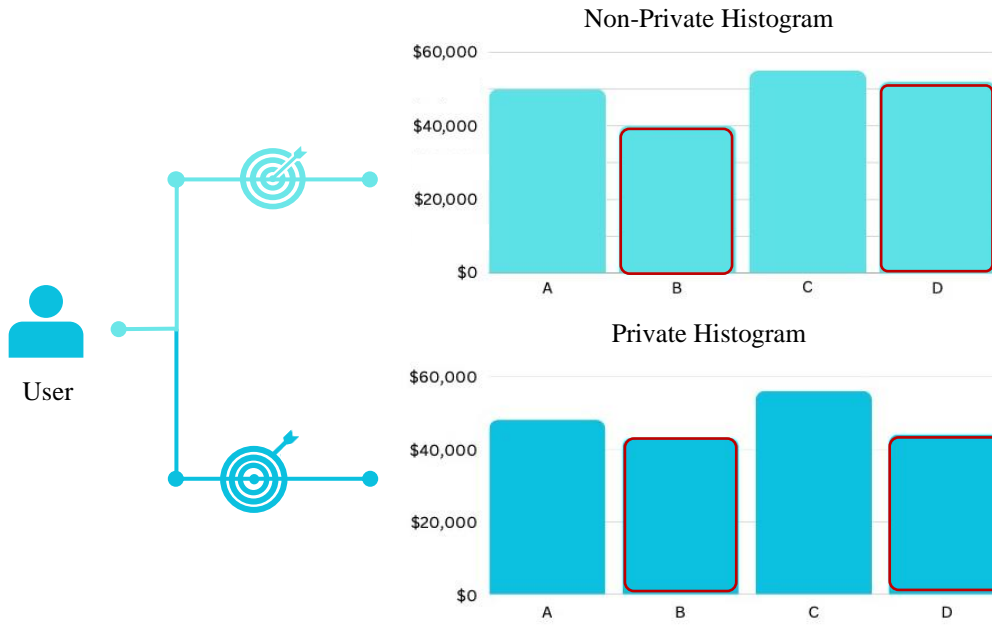


Figure 1: Effect of Differential Privacy on Evaluation Metrics.

Orthogonally, in the field of visual analytics, which is the focus of this work, noise-induced data perturbations in differentially private visualizations can alter visual patterns and impact the utility of the visualization (Zhang et al., 2021). On the other hand, lacking privacy protections increases the risk of disclosing private personal data. To visualize data with confidence regarding its privacy and utility, it is necessary to investigate and find a balanced approach. A real-world dataset of income levels in various regions is shown with bar charts in Figure 1. Changes that impact evaluation metrics are shown visually in the representation before and after differential privacy. The investigation of the privacy-utility trade-offs and effects of the differential privacy mechanism on data visualization will be based on this example.

In this work, we focus on studying the impact of data privacy on downstream data analytics, particularly histogram visualizations. We seek to provide insights that facilitate informed decision-making and raise the general efficacy of data visualization techniques by examining the ways in which privacy-preserving techniques impact the precision and utility of histograms. It's a crucial tool to comprehend the frequency and variability of numerical data and to visualize its distribution. The study also employs visualization strategies, like bar charts, to enhance comprehension and analysis. Bar charts allow analysts to easily compare the values of various subcategories within each main category, facilitating the detection of similarities and differences within and between groups (Diaz et al., 2018). This helps with informed decision-making during the visualization process.

Additionally, the study uses novel metrics to define privacy-utility trade-offs and data variety, integrating privacy gain, Normalized Mutual Information (NMI), and Sum Squared Error (SSE). To allow users to indicate which level of privacy they prefer over utility, we propose a hybrid multi-objective metric that combines these three metrics. This innovative technique offers a fresh perspective on optimization and a deeper comprehension of the intricate relationships that exist between data utility and privacy protection during the visualization process. Our method improves the adaptability and efficiency of privacy-preserving data visualization techniques by enabling customized preferences.

1.2 Problem Statement

This study aims to investigate the difficult trade-off between demonstrating the value of data visualization techniques and protecting individuals' privacy. The strict regulations enforced by the General Data Protection Regulation (GDPR) in the European Union (European Parliament and Council, 2016), the Health Insurance Portability and Accountability Act (HIPAA) in the United States (U.S. Department of Health and Human Services, 1996), and the UAE Federal Data Protection Law No. 45 of 2021 (UAE Government, 2021) have made strong privacy protection mandatory. This raises a fundamental question: Is it possible to maintain the confidentiality of sensitive data and still obtain valuable insights for research and decision-making?

The growing popularity of data-driven approaches in many different industries requires the development of practical solutions for safeguarding personal information. Therefore, by introducing noise, which can make an impact on the visualizations' accuracy, privacy techniques like differential privacy provide a way to safeguard individuals' sensitive data. Nonetheless, navigating this obstacle is still a challenging endeavor.

1.2.1 Research Questions

Based on previous motivation and concerns, we must discover solutions to the following research questions:

- RQ1: How do differential privacy frameworks impact individual privacy in data visualizations?
- RQ2: What are the trade-offs between preserving privacy and the accuracy of the visual representation in data visualization techniques?
- RQ3: How much do privacy-enhancing techniques affect the utility and precision of visualizations?
- RQ4: How can metrics be used to evaluate the trade-off between visualization quality and strong data privacy safeguards efficiently?

1.3 Research Objectives and Contributions

1. RQ1: Investigate the effectiveness of tuning Laplace noise parameters to accommodate different privacy requirements.
2. RQ1: Study the impact of various privacy parameter settings on data usability and privacy assurance, focusing on sensitive attributes.
3. RQ2: Analyze the impact of different privacy parameter settings that affect the utility of data visualization. Compare original and perturbed data visualizations to find the optimal balance between privacy and data utility.
4. RQ3: Assess the impact of the hybrid multi-objective metric on privacy protection and data utility trade-offs.

5. RQ4: Conduct a comprehensive analysis and mapping of the relationship between privacy settings and data utility, using empirical and estimated methods to determine the optimal privacy settings and ensure robust privacy protection.

The following summarizes the contributions considering the provided research questions:

- Examine the effectiveness of tuning Laplace noise parameters to meet different privacy requirements.
- Examine the effects of various privacy parameter settings on data usability and privacy protection and the utility of data visualization.
- Evaluate the impact of the hybrid multi-objective metric on balancing privacy protection and data utility.
- Conduct a comprehensive analysis of the relationship between privacy settings and data utility, identifying the optimal settings to ensure robust privacy protection and utility.

Chapter 2: Related Work

There has been a notable increase in interest and research in the field of privacy-preserving data analytics, especially in differential privacy, in recent years. This review explores both classic literature and recent scientific works with historical and modern viewpoints. Examining these sources, the review aims to clarify the development of differential privacy, evaluate the scope of current privacy-preserving methods, and identify new issues and trends that will influence the field's future developments.

A key idea in mitigating the possibility of unintentional disclosure of personal data when paired with other data is differential privacy. Many data privacy laws have been developed in response to growing concerns about data sharing to reduce the associated risks. An approach to measuring the degree of privacy protection provided by privacy-preserving data analysis techniques is provided by Differential Privacy. It aims to prevent the release of private data about specific individuals from a dataset (Bhattacharjee, 2020). The differential privacy approach has been extensively researched and used in a variety of fields, such as the social sciences, finance, and healthcare.

2.1 Foundations and Practical Applications of Differential Privacy

The idea explained in 'The Algorithmic Foundations of Differential Privacy' by Dwork and Roth (2014) emphasizes the creation of algorithms that can identify individual records in a dataset without making any specific inferences from this data. The theory of differential privacy, algorithmic implementation techniques, and practical applications are all thoroughly explored in this significant work, which is a cornerstone in the field of privacy-preserving data analysis. The contributions of Dwork and Roth have made it possible for researchers to study privacy-preserving algorithms from both theoretical and practical angles, which has aided in the development of new methods for addressing privacy concerns as well as advancements in privacy protection (Lindell, Y. and Pinkas, B., 2000).

Xiao and Tao investigated tailored approaches to privacy preservation in their 2015 paper "Personalized Privacy Preservation Techniques" (Xiao & Tao, 2006). They support the development of privacy-preserving methods that are tailored to the preferences and

privacy policies of everyone. The authors provide insights into the creation of privacy-conscious systems that can satisfy a range of user needs by integrating differential privacy with the specifications of recommendation systems for personalized data analysis.

The field of differential privacy has progressed, resulting in the creation of models and algorithms that analyze private data while maintaining data utility. These efforts encompass a range of strategies, including the Laplace mechanism (which focuses on controlled noise addition while maintaining established privacy guarantees) and Warner's Randomized Response (1965), as well as randomized mechanisms developed by McSherry and Talwar (2007).

The trade-offs between data utility and privacy have been examined in a number of studies. Federated Learning is a method that was presented by McMahan et al. (2017) and is intended to train machine learning models on decentralized data while maintaining privacy. They draw attention to the harmony that exists between privacy guarantees and model performance. The trade-offs between data utility and privacy in differentially private algorithms were also measured and optimized in a different study by Thakkar et al. (2021). The aforementioned studies highlight the significance of devising techniques that preserve data utility while guaranteeing confidentiality.

The Laplace mechanism is one method for differential privacy that is frequently employed. The Laplace process entails incorporating noise into the data from a Laplace distribution, which is defined by its scale parameter, which is based on the data's sensitivity and privacy parameter, ϵ (epsilon). A single data point's ability to affect the result is measured by sensitivity. The Laplace distribution's scale parameter is computed by dividing the sensitivity by ϵ . This method preserves the anonymity of the dataset's participants while enabling insightful statistical analysis (Dwork et al., 2006).

Using metrics to measure the trade-offs is a crucial part of striking a balance between privacy and utility. In this context, privacy gain and Normalized Mutual Information (NMI) metrics are important. The original and perturbed datasets are compared using the Normalized Difference Index (NMI), which gives information about how much of the original data structure remains after perturbation (Vinh, Epps, & Bailey, 2010). It is especially helpful for evaluating tasks involving clustering and classification,

where the data's structure is essential. NMI is a normalized metric with a range of 0 to 1, where 1 means that the original data's exact structure has been preserved in the perturbed data. Instead, privacy gain measures how much privacy protection has improved because of disturbance, indicating a lower chance of re-identification (Dwork et al., 2006). Comparing the privacy risks before and after applying the noise allows one to calculate the privacy gain. This metric helps to balance the trade-off between data utility and privacy protection by offering a quantitative assessment of the improvement in privacy (Dwork et al., 2006). These metrics assess the efficacy of privacy-preserving methods and their influence on data utility, allowing researchers to make well-informed decisions. Researchers can achieve a balanced approach to data privacy and utility by utilizing NMI and privacy gain. This ensures that data is protected from prying eyes while still being useful for analysis.

Table 1: Differentially Private Ages - Before and After Noise

Before	After
63	61.2
68	70.1
70	69.5
65	66.3
72	69.0

One of the core ideas of differential privacy is to provide a measurable indicator of privacy protection. By ensuring that a single data point cannot materially change the results of an analysis, this method helps to prevent the identification of individuals within a dataset (Dwork, 2006). Table 1 provides an example dataset with ages before and after the application of 2.5-parameter Laplace noise to demonstrate this idea. The following is the presentation of the resulting Laplace noise values for each age: 3.0, -0.5, 1.3, 2.1, and -1.8. After noise was applied, the ages measured were 61.2, 70.1, 69.5, 66.3, and 69.0. Privacy measures in the form of differentials are added to the dataset to ensure the confidentiality of individuals is maintained uncompromised while the overall utility of the data remains highly intact (Johnson, 2013).

2.2 Visualization Challenges in Privacy-Preserving Data Analysis

It isn't easy to comprehend data and share its insights without visualization tools. Plotting raw data or combined statistical measures is the standard method used to find patterns and correlations. However, visualizing data in the context of privacy presents different challenges. Methods like aggregation, generalization, and imputation are frequently employed to preserve insights while sanitizing and safeguarding sensitive data.

To safeguard sensitive data while still offering practical visual representations of the data, privacy-preserving data visualizations employ a variety of strategies. These methods include adding random stochastic noise, substituting centroids for individual-level observations, and performing anonymization procedures like k-anonymization (Zhou et al., 2023). Understanding the relationship between data-sharing policies and privacy parameters, striking a balance between privacy and utility, and spreading privacy-aware information all depend heavily on visualization (Dasgupta et al., 2013).

Histograms are essential tools for displaying data distributions but using them in situations where privacy is a concern calls for extra care. The requirement to protect individual data points must be balanced with the need to accurately represent the underlying data distribution in privacy-preserving histograms. To create privacy-preserving histograms, it is common practice to first add noise to the original data before visualizing it. For this, the Laplace mechanism is frequently employed, in which noise from a Laplace distribution is added to the data values prior to binning. The privacy parameter, ϵ , and the sensitivity of the query control the amount of noise (Dwork et al., 2006). This method protects people's privacy while enabling the histogram to offer insightful information about the distribution of the data.

In the context of a histogram, the sensitivity of a query is usually defined as the maximum amount that the data of a single individual can change the count in any bin. To ensure that the noise is enough to mask individual contributions without unduly distorting the overall data distribution, the scale of the noise added is proportional to this sensitivity divided by ϵ (Dwork et al., 2006).

Histograms that protect privacy make it possible to share data visualizations securely without disclosing private information. This is especially helpful in industries like healthcare, where sharing data can improve treatment and research without jeopardizing patient privacy. Histograms that adhere to strict privacy standards can yield actionable insights by adding controlled noise (Cormode et al., 2018).

2.3 Evaluation Metrics and Methodologies

There have been notable developments in privacy-preserving data analysis in addition to the previously mentioned contributions, particularly about visualization methods. In a comprehensive study, Zhang et al. investigated how participants' performance in visual data analysis was affected by different privacy levels, tasks, and types of visualization within the framework of differential privacy. They discovered that summary tasks were more resilient to higher noise levels when they conducted a comparison between summary tasks and value tasks. In addition, they presented a collection of measures for assessing data distribution before noise injection, which allowed analysts to manage their privacy budgets more effectively by reducing perceptual errors and enhancing the differential privacy model (Zhang et al., 2021).

In the context of digital privacy, assessing privacy metrics is essential to comprehending the trade-offs between NMI and privacy gain. Careful consideration of these metrics is necessary in a variety of scenarios, such as data sharing and information transfer, to guarantee that privacy is preserved without sacrificing the value of the data. New evaluation metrics grounded in information theory have been introduced in recent advances. These metrics offer a more nuanced understanding of how various factors, including the type of data being shared and the level of trust between parties, impact privacy. Metrics that consider the dynamic nature of trust variations, for instance, can offer insights into how privacy changes over time and enable more flexible privacy-preserving strategies (Wagner & Eckhoff, 2019).

Evaluating the degree to which privacy metrics safeguard user profiles is a further field of focus for privacy-improving technologies, especially in personalized information systems. Through the development of metrics that can measure the effect of privacy-preserving measures on the relevance and accuracy of personalized services, these

technologies seek to address the natural conflict between privacy and customization. All things considered, developing and improving privacy metrics is crucial to tackling the complex issues related to digital privacy. Researchers as well as developers can create more efficient and privacy-aware information systems by assessing privacy gain, NMI and trust dynamics (Wagner & Eckhoff, 2019).

Normalization is an important preprocessing step in the analysis of data and machine learning that unifies the range of different variables. It guarantees that every feature makes an equal contribution to the analysis and keeps features with bigger scales from taking priority over smaller-scale features. Scaling by the sum is a popular normalization method that uses the following formula to scale the data to a fixed range, typically between 0 and 1 (Majeed & Lee, 2021). Nevertheless, the normalization discussion is not a direct metric for evaluating privacy-preserving data analysis and will be covered in detail in the methodology section.

Normalized Mutual Information (NMI) and Privacy Gain: In privacy-preserving data analysis, NMI and Privacy Gain are essential evaluation metrics. According to Dwork et al. (2006), privacy gain quantifies the enhancement in privacy protection brought about by the disturbance and shows a decreased likelihood of re-identification. This measure ensures that the risk of exposing individual data points is kept to a minimum by evaluating how much privacy has been improved through techniques. However, NMI provides information about the percentage of the original data structure that remains after perturbation by comparing the perturbed and original datasets. According to Vinh, Epps, and Bailey (2010), NMI is especially helpful for evaluating tasks involving clustering and classification, where the data's structure is essential. Researchers can assess the effectiveness of privacy-preserving techniques and their impact on data utility by using NMI and privacy gain. This enables a balanced approach to data privacy and utility.

In the article "MuVE: Efficient Multi-Objective View Recommendation for Visual Data Exploration," the authors talk about how data reduction affects how original non-binned aggregate values are estimated. They observe that approximation errors in estimating the original values rise with a decrease in the number of bins. They use the Sum Squared Error (SSE), a commonly used metric for assessing accuracy in such

approximations, to quantify this error (Ehsan, Sharaf, & Chrysanthis, 2016). Also, In the paper "Histograms and Wavelets on Probabilistic Data," the authors talk about how to create wavelet and histogram synopses of probabilistic data using the Sum Squared Error (SSE) metric. SSE is a basic metric that's used to measure how accurate an approximation or estimation is, especially when it comes to data summarization methods like wavelets and histograms. By minimizing the squared differences between estimated and actual values, the SSE objective seeks to provide a numerical indicator of the approximation's quality (Cormode & Garofalakis, 2010). The quality of selectivity estimates in histograms is assessed by the authors of the paper "Optimal Histograms with Quality Guarantee" using the Sum Squared Error (SSE) as a crucial metric. Selectivity estimates depend on SSE to detect outliers and mark them as inappropriate for further processing. Therefore, SSE is used in selectivity estimation. To increase the precision of each query estimate, the authors suggest improving histograms to offer quality guarantees on selectivity estimates for equality and range queries (Jagadish et al., 1998).

Two other metrics, Average SSE and Root Sum of Squares (RSS) are frequently used in addition to the Sum Squared Error (SSE) to evaluate the accuracy and quality of data approximate values and summaries. The average SSE, which produces a mean error measure by normalizing the SSE by the number of observations, is especially helpful when comparing the precision of various models or summaries over datasets of various sizes. It is used in situations where averaging the total error is necessary to determine the error per data point, providing a more comprehensive view of the quality of the approximation per unit (Zhou & Troyanskaya, 2007).

Root Sum of Squares (RSS), on the other hand, is the square root of SSE and provides a more comprehensible metric by returning the error measure to the original data units. This makes it easier to relate errors to the actual data values by helping to visualize and comprehend the magnitude of errors. When utilizing probabilistic data summarization and approximation techniques, the RSS is frequently combined with other statistical measures to offer a comprehensive understanding of the error distribution (Tukey, 1977). These metrics are essential for guaranteeing the accuracy and consistency of data approximations and summaries, which facilitates efficient data exploration and decision-making procedures.

To address privacy concerns, other strategies such as k-anonymity, l-diversity, and t-closeness have also been proposed in addition to the methods mentioned above. K-anonymity makes re-identification more difficult by ensuring that everyone in a dataset cannot be distinguished from at least $k-1$ other individuals based on certain attributes. By guaranteeing that sensitive attributes have at least l "well-represented" values, L-diversity extends k-anonymity and thwarts homogeneity attacks. By guaranteeing that the distribution of a sensitive attribute in any equivalency class is close to its distribution in the entire dataset, T-closeness improves upon these ideas and prevents attribute disclosure (Machanavajjhala et al., 2007).

Differential privacy has been introduced in federated learning scenarios in more recent works. Federated learning improves privacy by enabling several parties to work together to jointly train machine learning models without exchanging raw data. To provide an extra degree of privacy protection, this technique has been combined with differential privacy, guaranteeing that individual participant contributions remain hidden even when the model is updated. Particularly in applications like medical data analysis, where data sharing is frequently restricted owing to privacy concerns, these methods have demonstrated promise in striking a balance between privacy and utility (McMahan et al., 2017).

2.4 Navigating the Conflict: Privacy Needs vs. Utility

Many studies have investigated topics like data sharing's usefulness and privacy over time. The difficulty is in implementing the right modifications into place to protect privacy rights without sacrificing the dataset's analytical utility. Differential privacy is one privacy-preserving technique, but it may introduce noise and distortion, which could affect the accuracy and usefulness of downstream analysis.

There is often a lack of research on evaluating the trade-offs between privacy and utility and creating metrics to measure these trade-offs. Although differential privacy provides a common mechanism for formal privacy guarantees, there is still work to be done in creating thorough assessment metrics that consider a system's utility and privacy aspects (Ivanova, 2022). Similarly, research questions for a particular subject universe may be addressed when respondents are excluded, and researchers concentrate only on

domains or data types. However, it becomes difficult to apply these findings to a different universe with different conditions.

Several methods are used to assess how well privacy and utility are balanced in data perturbation. A possible approach is to evaluate the effectiveness of different perturbation techniques in terms of attaining the required privacy and utility levels by comparing them (Roman, 2023). An alternative approach for attaining differential privacy (DP) involves adjusting the random noise incorporated into data values while considering the consequences for dataset availability (Ma et al., 2023). Furthermore, by obtaining perturbation from other samples in the dataset, stochastic perturbation techniques can sanitize datasets. These techniques provide a sophisticated understanding of how to balance privacy and utility when dealing with data perturbation.

In conclusion, managing the balance between the need for privacy and the usefulness of data needs a thorough comprehension of the many perturbation methods, thorough assessment metrics, and effective visualization techniques. Researchers and practitioners can make sure that data is still useful for analysis while protecting individual privacy by carefully balancing these elements. Table 2 highlights the main contributions made in privacy-preserving data analysis and compares them, emphasizing their respective areas of focus and conclusions.

Table 2: Summary Table of Existing Work and Their Comparison.

Author(s)	Year	Contribution	Key Findings
Dwork & Roth	2014	Algorithmic Foundations of Differential Privacy	Established theoretical and practical aspects of differential privacy
Xiao & Tao	2006	Personalized Privacy Preservation Techniques	Advocated for privacy-preserving methods tailored to individual preferences
McSherry & Talwar	2007	Randomized Mechanisms	Developed randomized mechanisms for differential privacy
Vinh, Epps, & Bailey	2010	Normalized Mutual Information (NMI)	Introduced NMI for assessing the preservation of data structure after perturbation
Zhang et al.	2021	Impact of Privacy Levels on Visual Data Analysis	Found summary tasks more resilient to higher noise levels
Wagner & Eckhoff	2019	Information Theory-Based Privacy Metrics	Offered nuanced insights into the impact of data sharing and trust dynamics on privacy
Ehsan, Sharaf, & Chrysanthis	2016	Data Reduction and Sum Squared Error (SSE)	Quantified errors in estimating original values due to data reduction using SSE
Cormode & Garofalakis	2010	Histograms and Wavelets on Probabilistic Data	Discussed creating wavelet and histogram synopses of probabilistic data using SSE
Jagadish et al.	1998	Optimal Histograms with Quality Guarantee	Improved histograms to offer quality guarantees on selectivity estimates using SSE
Tukey	1977	Exploratory Data Analysis	Introduced Root Sum of Squares (RSS) for better understanding of error magnitudes
Ivanova	2022	Evaluation of Trade-offs Between Privacy and Utility	Highlighted the need for comprehensive assessment metrics that consider both privacy and utility
Roman	2023	Effectiveness of Different Perturbation Techniques	Evaluated perturbation techniques in terms of privacy and utility balance
Ma et al.	2023	Stochastic Perturbation Techniques	Proposed methods for balancing privacy and utility using stochastic perturbation

Chapter 3: Methods

This chapter describes the research methodology for analyzing the trade-offs between data utility and privacy protections in visualization techniques. To safeguard critical information, a descriptive strategy was used along with differential privacy strategies, particularly the Laplace mechanism. Our methodology seeks to close gaps in the literature and capture the subtleties of privacy-preserving techniques while offering a thorough understanding of how differential privacy impacts data visualization.

3.1 Overview

To examine the trade-offs between privacy protections and data utility in visualization techniques, the dissertation research methodology used a descriptive approach and included a differentially private special technique (Erlingsson, Pihur, & Korolova, 2014). The descriptive design was selected in order to offer an extensive understanding of differential privacy strategies and how they affect data visualization. This methodology makes sure that the study fully characterizes and examines the phenomenon of privacy-preserving data visualization, filling in the knowledge gaps in the literature and capturing the nuanced aspects of how differentiating privacy techniques affect the results of visualization (Zhang et al., 2021).

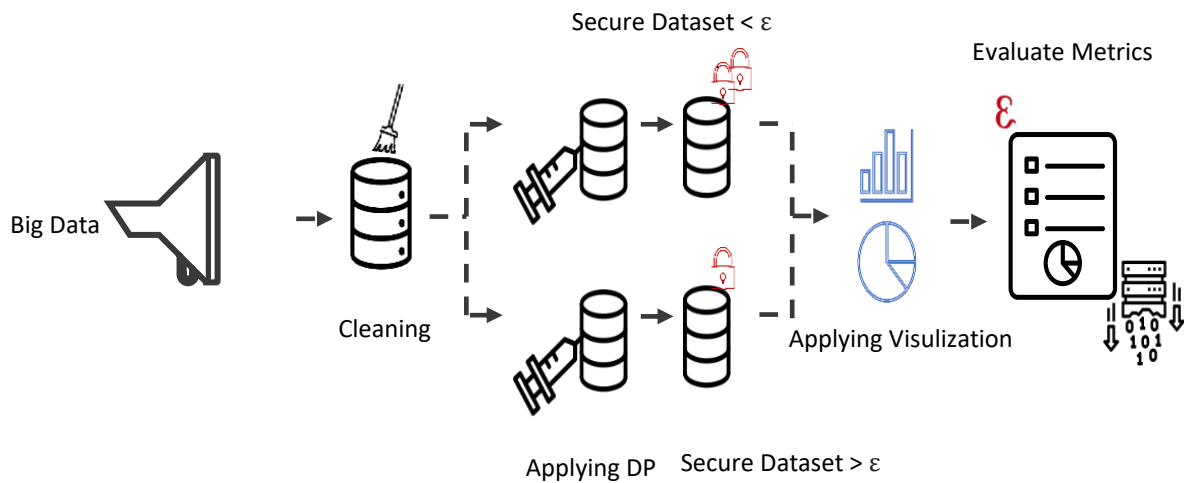


Figure 2: Overview of the Research.

In this study, as illustrated in Figure 2, the dataset was chosen and cleansed; differential privacy was applied by incorporating Laplace noise into the data with different epsilon values before producing visualizations. This strategy aimed to protect private or sensitive data from unintentional exposure using visualizations. Aggregated bar charts were used for data visualization, which is a useful way to show summary statistics or combined values from several attributes. These graphs facilitated the understanding of patterns across a variety of variables and allowed for simple comparisons. Furthermore, histograms were used to group numerical data into bins and display the frequency or count of data points within particular intervals. In addition, the data's usefulness and privacy were evaluated by different metrics. By calculating the squared differences between the original and estimated values, the Sum Squared Error (SSE) was used to quantify the accuracy of data approximations and provide a numerical indicator of the caliber of data summarization techniques. The improvement in privacy protection brought about by disturbance was measured as privacy gain, which reflected the lower risk of re-identification. To determine how much of the original data structure remained after perturbation, the similarity between the original and perturbed datasets was measured using Normalized Mutual Information (NMI). After comparing the results of these metrics, a hybrid method that integrated both was developed to gauge the overall privacy and usefulness of the data.

3.2 Applying Data Privacy

To address our research questions, we experimented on a dataset. We started by using a popular differential privacy algorithm called the Laplace Mechanism. It's a standard method for achieving differential privacy by adding carefully calibrated noise to data values, thereby altering the original data. Specifically, the noise is generated from a Laplace distribution with a mean of zero and a scale factor determined by the epsilon parameter and sensitivity of the data. This process results in a noise value being added to each data point. The Laplace Mechanism, where ϵ is the privacy parameter, adds noise $\text{Lap}(\Delta f/\epsilon)$ to the result of f , given a sensitivity Δf . The differential privacy of f 's output is ensured by this addition of noise. Here's a formula to indicate how Laplace noise can be added to each value in a dataset X :

The Laplace distribution is mathematically formulated as:

$$Lap(x | \mu, b) = \frac{1}{2b} \exp \left(- \frac{|x - \mu|}{b} \right) \quad (1)$$

Where μ is the mean and b is the scale parameter.

Given a dataset $X = \{x_1, x_2, \dots, x_n\}$ with a sensitivity parameter ΔX , the Laplace noise η for each value x_i can be generated as follows:

$$\eta_i = Lap(0, \frac{\Delta X}{\epsilon}) \quad (2)$$

The perturbed dataset X' is then obtained by adding the noise to the original dataset:

$$X' = x_i + \eta_i \quad (3)$$

Where ϵ is the privacy parameter, and $\Delta X/\epsilon$ is the scale parameter of the Laplace distribution.

In our study, we applied Laplace noise to all values in dataset X . We can now better examine the trade-offs between privacy and utility in real-world datasets and understand the impact of privacy-preserving mechanisms on data analysis due to this procedure. When adding Laplace noise to dataset attributes, it is important to remain careful and professional. Usually, the sensitivity of the data determines the scale parameter of a Laplace distribution with a mean of 0. As the maximum possible change in output that could arise from adding or subtracting a single value, we choose a sensitivity value of 1 in our situation.

Laplace noise incorporation into dataset attributes requires a careful and expert approach. Usually, this noise applies from a Laplace distribution with a mean of 0 and a scale parameter that depends on how sensitive the data are. We use a sensitivity value of 1 in our scenario, which represents the maximum possible change in output that could arise from adding or subtracting a single value.

To demonstrate, let's consider a dataset that includes individuals' ages that we would like to visualize using a bar chart. The sensitivity of this operation is set to 1, meaning that the addition or removal of a single individual from the dataset will identically impact the counts in one of the histogram bins by a factor of 1. Consequently, each histogram bin receives random noise samples from $\text{Laplace}(1/\epsilon)$ using the Laplace mechanism, producing a noisy histogram.

This technique ensures insightful analysis and visualization of the dataset while maintaining the privacy of the data. By following these steps, we can implement Laplace noise to our data effectively and achieve the desired privacy and utility balance by following these procedures (Zhang et al., 2021).

To promise the reproducibility of our findings, we set a seed value of 42 for the random number generator used in producing the Laplace noise. This indicates that the same set of randomly selected numbers is generated each time the Laplace Mechanism is applied, enabling precise reproduction of the results we obtained.

One challenge that we faced was handling noise, that might trigger the perturbed data values to fall outside the original data's valid range. For example, if the original dataset contains ages ranging from 17 to 99 years, adding noise could produce values that are negative or exceed 99. To tackle this problem, we implemented a method to check the range of the perturbed values. If a perturbed value falls outside the valid range, we discard it and reproduce new noise values until we find a valid value within the range. This guarantees the perturbed data's authenticity and consistency with the original dataset. We seek to balance privacy and data utility in our study by employing the Laplace Mechanism with careful consideration of the noise parameters and handling of out-of-range values.

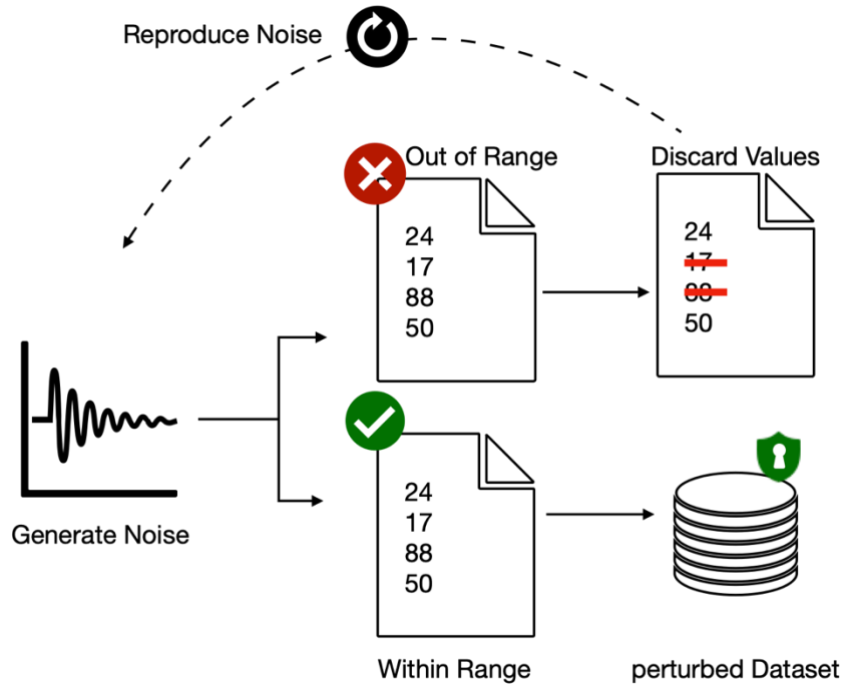


Figure 3: Handling Out-of-Range Noise Values in Laplace Perturbation.

Figure 3 illustrates the procedure for managing noise values that are outside of the acceptable range when applying Laplace noise to a dataset. It starts by producing Laplace noise and then determines whether the perturbed value is inside the acceptable range. The procedure moves on to the next step if the value is within the range. If not, new Laplace noise is produced, and the out-of-range value is discarded. Until a valid number within the range is found, this cycle of creating noise, examining the range, and then discarding or regenerating noise is repeated. Using this method preserves the integrity and authenticity of the perturbed data by making sure that its values fall within the valid range of the original dataset.

3.3 Privacy Metrics

Evaluation of privacy mechanisms' effectiveness in protecting sensitive data while preserving data utility is crucial for privacy-preserving data analysis. Normalized Mutual Information (NMI) and Privacy Gain are two important metrics used in this assessment.

Privacy Gain measures the enhancement of privacy protection generated by data disturbance, indicating the decreased probability of re-identifying individuals within a

dataset. This measure is essential for determining how well sensitive information is protected by privacy-preserving measures. Based on differential privacy principles, the idea of Privacy Gain makes sure that no single person's data has a substantial impact on the results of any analysis, which keeps private information from being revealed (Dwork et al., 2006). The formula for Privacy Gain (PG) is:

$$\text{Privacy Gain} = \max_i |x_i - \hat{x}_i| \quad (4)$$

Where x_i is the original value of data point i , and \hat{x}_i is the perturbed value of data point i , and \max_i denotes taking the maximum absolute difference over all data points.

Let's consider a simple dataset of annual incomes (in thousands of dirhams), the data will be perturbed using the Laplace noise mechanism. We'll then calculate the Privacy Gain to determine the extent to which privacy protection has improved. Let's consider that the original dataset is [50, 60, 75, 90, 110], and the perturbed dataset after applying the Laplace Noise becomes [55, 56.5, 77.3, 85.3, 116.1]. We compute the Privacy Gain by determining the differences for every pair: $|50 - 55| = 5.0$, $|60 - 56.5| = 3.5$, $|75 - 77.3| = 2.3$, $|90 - 85.3| = 4.7$, and $|110 - 116.1| = 6.1$.

The Privacy Gain is represented by the maximum difference of 6.1. The degree to which individual values have been modified to preserve privacy is reflected in this value, which represents the highest degree of deviation added to the data points. The Privacy Gain metric measures the amount to which the dataset's privacy has improved, guaranteeing that each individual data point is substantially altered to avoid re-identification while maintaining the data's overall analytical value.

The mutual dependence between the original and perturbed datasets is measured using the normalized mean value (NMI). It measures the degree to which the data structure is maintained following disturbance, which is significant for jobs requiring data integrity (Vinh, Epps, & Bailey, 2010). NMI values vary from 0 to 1, where a value of 1 guarantees high utility since the perturbed data maintains the exact structure of the original data.

Given two datasets XX (original data) and YY (perturbed data), the Mutual Information $I(X;Y)$ is defined as:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \quad (5)$$

Where $p(x, y)$ is the joint probability distribution of XX and YY , which measures the probability that a particular pair of values (x,y) occurs together, $p(x)p(y)$ is the marginal probability distribution of XX (original data), and $p(y)p(y)$ is the marginal probability distribution of YY (perturbed data). The interpretation where $x \in X$ is the values from the original dataset, $y \in Y$ is the values from the perturbed dataset, $p(x, y)$ is the probability that the pair (x, y) appears together, i.e., both the original value x and the perturbed value y , and $\log \left(\frac{p(x,y)}{p(x)p(y)} \right)$ measures how much knowing x reduces the uncertainty about y and vice versa.

Normalized Mutual Information (NMI) adjusts the MI to scale between 0 and 1, providing a normalized measure of mutual dependence. It is defined as:

$$NMI(X;Y) = \frac{2 \cdot I(X;Y)}{H(X) + H(Y)} \quad (6)$$

Where $I(X;Y)$ is the Mutual Information between X (original data) and Y (perturbed data) and $H(X)$ is the entropy of the original data:

$$H(X) = - \sum_{x \in X} p(x) \log p(x) \quad (7)$$

$H(Y)$ is the entropy of the perturbed data:

$$H(Y) = - \sum_{y \in Y} p(y) \log p(y) \quad (8)$$

Let's look at a small dataset of values to demonstrate how to calculate Normalized Mutual Information (NMI) between original and perturbed datasets. Let's say we have a dataset consisting of numbers [25, 30, 35, 40, 45]. After applying Laplace noise for privacy preservation, we obtain a perturbed dataset [26, 29, 37, 38, 44]. Initially, we separate the original and perturbed datasets into distinct bins using discretization. Assume that for this procedure, ten bins are used. The original dataset's discretized values could be [0, 1, 2, 3, 4], and for the perturbed dataset, they might be [0, 0, 2, 2, 4].

We then compute the discretized values' marginal probabilities for the original and perturbed datasets. For the original dataset, the probabilities are: $P(X = 0) = 1/5$, $P(X = 1) = 1/5$, $P(X = 2) = 1/5$, $P(X = 3) = 1/5$, and $P(X = 4) = 1/5$. For the perturbed dataset, the probabilities are: $P(Y = 0) = 2/5$, $P(Y = 2) = 2/5$, and $P(Y = 4) = 1/5$.

Next, we calculate the pairs of discretized values' joint probabilities. For our example, the joint probabilities are: $P(X = 0, Y = 0) = 1/5$, $P(X = 1, Y = 0) = 1/5$, $P(X = 2, Y = 2) = 1/5$, $P(X = 3, Y = 2) = 1/5$, and $P(X = 4, Y = 4) = 1/5$.

We apply the entropy formula to determine the entropies of the original and perturbed datasets:

$$H(X) = - \sum_{x \in X} p(x) \log p(x) \quad (9)$$

For the original dataset, the entropy is calculated as follows:

$$H(X) = -5 \times \frac{1}{5} \log \frac{1}{5} = \log 5 \approx 1.60944$$

For the perturbed dataset, the entropy is:

$$H(Y) = -\left(\frac{2}{5} \log \frac{2}{5} + \frac{2}{5} \log \frac{2}{5} + \frac{1}{5} \log \frac{1}{5}\right) \approx 1.05492$$

Next, we calculate the mutual information (MI) between the original and perturbed datasets using the formula:

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right)$$

For each pair (X, Y) :

$$I(0; 0) = \frac{1}{5} \log \left(\frac{\frac{1}{5}}{\frac{1}{5} \times \frac{2}{5}} \right) = 0.13863$$

$$I(1; 0) = \frac{1}{5} \log \left(\frac{\frac{1}{5}}{\frac{1}{5} \times \frac{2}{5}} \right) = 0.13863$$

$$I(2; 2) = \frac{1}{5} \log \left(\frac{\frac{1}{5}}{\frac{1}{5} \times \frac{2}{5}} \right) = 0.13863$$

$$I(3; 2) = \frac{1}{5} \log \left(\frac{\frac{1}{5}}{\frac{1}{5} \times \frac{2}{5}} \right) = 0.13863$$

$$I(4; 4) = \frac{1}{5} \log \left(\frac{\frac{1}{5}}{\frac{1}{5} \times \frac{1}{5}} \right) = 0.32189$$

Summing these values:

$$I(X; Y) = 0.13863 + 0.13863 + 0.13863 + 0.13863 + 0.32189 = 0.87641$$

Normalized Mutual Information (NMI):

$$\text{NMI}(X; Y) = \frac{2 \cdot I(X; Y)}{H(X) + H(Y)}$$

$$\text{NMI}(X; Y) = \frac{2 \cdot 0.87641}{1.60944 + 1.05492} \approx 0.658$$

Thus, the NMI for this example is approximately 0.658, indicating a moderate preservation of the original data structure after perturbation.

Therefore, in this example, an NMI value of roughly 0.658 indicates that 65.8% of the original data structure is retained in the perturbed data. This indicates a moderate utility level, indicating that a substantial portion of the original structure is retained in the perturbed data, which is important for tasks that depend on data integrity.

3.4 Generating Visualizations/Histograms

Visualizations are invaluable tools for data analysis, given their ability to shed light on data distributions and patterns. Accurately and meaningfully visualizing sensitive data while maintaining privacy is a major challenge in privacy-preserving settings. Bar charts and histograms were the main visualization tools used in this study.

Numerical data can be effectively displayed as a frequency distribution using histograms, which count the number of data points within each bin and divide the data into bins. To ensure privacy, we first applied Laplace noise to the original data values in our study before creating histograms. Individual data points are protected by the Laplace mechanism, which introduces noise based on the sensitivity of the data and the selected epsilon value. To create the histogram, this process entails perturbing each data point with noise and then binning the resulting data. We can visualize data distributions using this method without losing personal privacy.

Histogram bin width selection is important since it has a big impact on the accuracy and readability of the visualization. The shape and meaning of the histogram are influenced by the bin width, which also determines how the data is grouped. Overly broad bins may oversimplify the data, hiding significant patterns and details, while overly narrow bins may overcomplicate it, adding noise and making it challenging to identify significant trends.

In this study, the bin width selection procedure was automated using Sturges' rule. The number of bins in a histogram can be found using the widely used Sturges' rule

method. By determining the bin width using the sample size's logarithm, it offers a methodical and theoretically supported approach. The definition of the rule is:

$$k = 1 + \log_2(n) \quad (10)$$

where n is the total number of observations in the dataset and k is the number of bins. Sturges' rule is used to calculate the number of bins, guaranteeing that the bin width is suitable for the dataset size and standardizing the assessment of trade-offs between privacy and utility across different data attributes and epsilon values.

Choosing a suitable bin width becomes even more important in privacy-preserving histograms. A well-selected bin width maintains a balance between privacy and utility by ensuring that the additional Laplace noise does not significantly alter the data distribution. The efficacy of privacy-preserving visualizations depends on the careful selection of bin widths appropriate for the size and distribution properties of the data. By striking a balance between maintaining the overall data utility and safeguarding individual privacy, this deliberate selection makes it possible to conduct accurate and insightful data analysis.

For this study, histograms were selected as the main visual tool for many reasons. They offer a simple method for visualizing the distribution of numerical data, exposing central patterns, variability, and outliers that are crucial for comprehending the data as well as underlying data distributions. The histogram visualization method works well with the addition of Laplace noise to individual data points because it accommodates noise introduced for privacy preservation while still giving insightful results. Histograms can also be used to analyze a variety of datasets with varying properties because they can be adjusted to different levels of granularity through bin width changes.

High-dimensional data is better represented by other visualization tools, although histograms are useful for representing one-dimensional data distributions. Heatmaps use color intensity to display data patterns and correlations in matrices, while scatter plots can be used to show relationships between several variables. Simpler visualization is made possible by dimensionality reduction techniques like PCA, which lower data dimensions while maintaining structure. Therefore, high-dimensional data requires more complex

visualization techniques, even though histograms are an effective tool for visualizing one-dimensional data distributions. The features of the data and the insights that are required will determine which visualization tools are best. While preserving privacy, combining various visualization techniques can offer a thorough understanding of high-dimensional datasets.

3.5 Accuracy/SSE Metrics

We employed several metrics, such as the Root Sum of Squares (RSS), Average SSE, and Sum Squared Error (SSE), to assess the precision and caliber of data approximations in privacy-preserving scenarios. These measures offer a thorough grasp of how well data utility is maintained by the privacy-preserving methods.

An extensively used metric for evaluating approximation accuracy is the Sum Squared Error (SSE). By calculating the squared differences between the estimated and actual values, it quantifies the total error and offers a numerical indication of how good the approximation is. SSE is especially helpful for assessing data summarization techniques such as wavelets and histograms. More accurate approximations are indicated by lower SSE values, which aid in our comprehension of how privacy-preserving methods affect data utility. The formula for SSE is:

$$SSE = \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (11)$$

Where n is the total number of bins, y_i is the observed frequency of data points in the i th bin and \hat{y} is the expected frequency of data points in the i th bin based on the perturbed data.

The mean error measure that results from normalizing the SSE by the quantity of observations is called the average SSE. This measure is helpful for evaluating the accuracy of various models or summaries over a range of dataset sizes. A balanced evaluation of privacy and utility is made possible by the average SSE, which contributes to a more thorough picture of the approximation quality per data point.

The square root of SSE is called Root Sum of Squares (RSS), and it returns the error measure to the original data units. The visualization and understanding of error magnitudes are facilitated by this metric, which makes it simpler to relate errors to the actual data values. A comprehensive grasp of error distribution in probabilistic data summarization and approximation techniques can be obtained by combining RSS with other statistical measures.

Normalization is an important preprocessing step in the analysis of data and machine learning that unifies the range of different variables. It guarantees that every feature makes an equal contribution to the analysis and keeps features with bigger scales from taking priority over smaller-scale features (Majeed & Lee, 2021). Scaling by the sum is a popular normalization method that uses the following formula to scale the data to a fixed range, typically between 0 and 1:

$$\text{Normalized value} = \frac{\text{Original value}}{\text{Sum of all values}} \quad (12)$$

where Original value is the value of the variable before normalization, and Sum of all values is the sum of all values of the variable in the dataset. This equation ensures that the normalized values fall within the specified range. By removing the impact of scale changes, normalization confirms an equal range of comparisons. This is needed when comparing variables with different scales or units as we did in this research.

The original and perturbed data values were normalized in our analysis before being plotted in histograms and the frequencies (or "counts") in each bin were determined. By using this method, we could be sure that our SSE and other accuracy metrics were calculated on a similar scale, which allowed us to draw useful conclusions about the trade-offs between privacy and data utility.

Think about a dataset, for instance, that compares ages before and after Laplace noise was added to protect privacy. The perturbed ages following the addition of Laplace noise with $\epsilon = 1.0$ could be [22, 24, 32, 34, 38], while the original ages could be [20, 25, 30, 35, 40].

$$\text{Sum of original values} = 20 + 25 + 30 + 35 + 40 = 150$$

$$\text{Sum of perturbed values} = 22 + 24 + 32 + 34 + 38 = 150$$

Normalized original ages:

$$[\frac{20}{150} + \frac{25}{150} + \frac{30}{150} + \frac{35}{150} + \frac{40}{150}] = [0.1333, 0.1667, 0.2000, 0.2333, 0.2667]$$

Normalized perturbed ages:

$$[\frac{22}{150} + \frac{24}{150} + \frac{32}{150} + \frac{34}{150} + \frac{38}{150}] = [0.1467, 0.1600, 0.2133, 0.2267, 0.2533]$$

The Sum Squared Error (SSE) between the normalized original and perturbed ages is then computed:

$$SSE = \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

$$SSE = (0.1467 - 0.1333)^2 + (0.1600 - 0.1667)^2 + (0.2133 - 0.2000)^2 + (0.2267 - 0.2333)^2 + (0.2533 - 0.2667)^2 = 0.00062446$$

Next, we calculate the Average SSE, which is the SSE divided by the number of observations:

$$Average\ SSE = \frac{SSE}{number\ of\ observations} \quad (13)$$

$$Average\ SSE = \frac{0.00062446}{5} = 0.00012489$$

The Root Sum of Squares (RSS), or square root of the SSE, is then computed:

$$RSS = \sqrt{SSE} = \sqrt{0.00012489} = 0.0250$$

To evaluate the precision and consistency of data approximations and summaries, these metrics and normalization techniques are essential. By guaranteeing that the data stays valuable while safeguarding individual privacy, they enable effective data exploration and decision-making processes. We can assess privacy-preserving techniques' efficacy and make informed decisions about how to balance privacy and utility in data analysis by carefully implementing these metrics.

3.6 Hybrid Metric

We developed two hybrid metrics that combine privacy and utility measures to achieve a balanced assessment of both. To ensure that privacy protection and data utility are sufficiently addressed, these composite metrics combine the strengths of individual metrics to provide a comprehensive evaluation.

The first hybrid metric aims to strike a balance between accuracy and privacy gain. This metric's objective is to maximize it because higher values indicate improved data utility and privacy protection. The following is how this hybrid metric is generated:

$$HM_{PG-Acc} = w_{utility} \times (1 - SSE) + w_{privacy} \times Privacy\ Gain$$

Where $w_{utility}$ is the weight assigned to the accuracy measure, $w_{privacy}$ represents the weight assigned to privacy gain, SSE is Sum Squared Error, PG is Privacy Gain and Acc is the accuracy from $1 - SSE$.

By changing the weights $w_{utility}$ and $w_{privacy}$, researchers can give priority to either accuracy or privacy, giving a flexible and nuanced evaluation of the associated trade-offs.

The second hybrid metric seeks to achieve balance between SSE and NMI (Normalized Mutual Information). This metric's objective is to minimize it because lower values indicate better privacy-preserving data utility preservation. The following is how this hybrid metric is produced:

$$HM_{NMI-SSE} = w_{utility} \times SSE + w_{privacy} \times NMI \quad (14)$$

Where $w_{utility}$ is the weight assigned to the SSE measure, $w_{privacy}$ represents the weight assigned to NMI , SSE is Sum Squared Error, NMI is the Normalized Mutual Information.

Consider the following example for clarification. Let us examine a dataset with the following values: 0.7 for Normalized Mutual Information (NMI), 0.02 for Sum Squared Error (SSE), and 0.8 for Privacy Gain (PG). The hybrid metrics can be computed as follows if the weights for privacy gain $w_{privacy} = 0.5$ and the weight of the Accuracy $w_{utility} = 0.5$. The hybrid metric for the Privacy Gain and Accuracy:

$$HM_{PG-Acc} = 0.5 \times (1 - 0.02) + 0.5 \times 0.8 = 0.89$$

The hybrid metric for the Privacy Gain and Accuracy:

$$HM_{NMI-SSE} = 0.5 \times 0.02 + 0.5 \times 0.7 = 0.36$$

The outcomes of these hybrid metrics show how successful the privacy-preserving methods are. For instance, a high score of 0.89 for the Hybrid Metric with Privacy Gain and Accuracy indicates that preserving privacy and obtaining high data utility have been well-balanced. This indicates that the perturbation applied to the data successfully maintains accuracy and privacy.

In the second scenario, a lower Hybrid Metric with NMI and SSE value of 0.36 indicates that the perturbation preserves the data utility while still maintaining the data's structure and obtaining a low SSE. This balance is critical for tasks that need to minimize approximation errors while preserving the data's privacy.

The total of the hybrid metric formulas' weights for utility $w_{utility}$ and $w_{privacy}$ must equal 1. This guarantees that utility and privacy are fairly considered in the combined metric, which is represented by the equation $w_{utility} + w_{privacy} = 1$. Also, it is necessary to normalize the Sum Squared Error (SSE) to a value between 0 and 1 to meaningfully combine it with other metrics. For Privacy Gain to be consistent with the term $(1 - SSE)$, it needs to be scaled appropriately. This guarantees a comparable contribution from each component to the hybrid metric. However, NMI is directly comparable to the normalized SSE because it naturally ranges from 0 to 1.

We can evaluate the trade-offs between privacy and utility because of these hybrid metrics. We can ensure that data analysis strikes a balance between privacy and utility by carefully applying these metrics, which will allow us to preserve individual privacy while maintaining the data's utility.

3.7 Optimizing Privacy Setting

Achieving privacy and utility balance requires optimizing the privacy parameter epsilon (ϵ). Epsilon regulates the amount of noise added to the data; higher utility but lower privacy is offered by smaller values of ϵ , while larger values offer both higher privacy and lower utility. The challenging aspect is figuring out the ideal ϵ value that best balances data utility preservation and sensitive information protection.

We tested various epsilon values and assessed the resulting privacy and utility metrics to identify the ideal epsilon value. Our goal was to determine the optimal value for privacy protection and data utility by examining the trade-offs between privacy gain, Normalized Mutual Information (NMI), and Sum Squared Error (SSE) for various epsilon settings.

There were several important steps in the optimization process. Initially, for every epsilon value, we calculated the hybrid metric. This hybrid metric provided a thorough analysis of every epsilon setting by integrating privacy and utility measures. Higher values denote better privacy protection and data utility, so the hybrid metric for privacy gain and accuracy (PG-Acc) was developed to maximize. Lower values indicate better preservation of data utility while maintaining privacy, so the hybrid metric for NMI and SSE was designed to minimize.

We used a quadratic curve fitting procedure to determine the ideal epsilon value. For every epsilon setting, this required fitting a quadratic curve to the total metric values and figuring out which epsilon value minimized the quadratic curve. The following steps can be used to explain the quadratic fitting process:

Initially, for every epsilon value, we calculated the hybrid metric. The hybrid metric values were then fitted with a quadratic curve. Ultimately, the epsilon value that minimized the quadratic curve was found. The relationship between epsilon and the hybrid metric could be precisely modeled by fitting a quadratic curve, which would enable us to determine the ideal epsilon value. This method makes sure that the epsilon value selected offers the best trade-off between safeguarding sensitive data and maintaining the data's usefulness.

We computed the optimal epsilon value automatically for each combination of privacy and utility weights. We systematically found the epsilon value that minimized or maximized the hybrid metric, depending on the objective, by fitting a quadratic curve to the hybrid metric values for each weight combination. Through this process, we were able to determine the ideal epsilon value for every unique weight combination, guaranteeing a customized trade-off between privacy and usefulness for various analytical purposes.

We analyzed and optimized the trade-offs between privacy and utility in privacy-preserving data visualizations in a methodical manner using these metrics and approaches. Researchers and industry professionals gained important insights from this optimization process, empowering them to decide on the right degree of privacy protection to implement without compromising the usability of their data. In order to ensure that the data is suitable for analysis while sufficiently protecting individual privacy, we were able to strike a significant balance between privacy gain and utility measures, for example, by identifying the optimal epsilon values through this process. This optimization guarantees that sensitive information can be analyzed efficiently without compromising it, while also improving the reliability of the privacy-preserving methods.

Chapter 4: Results and Discussions

This chapter explores the implications of the results obtained by using the methods that have been described. The results offer vital insights into the fine balance that exists between privacy protection and data utility within the framework of differentially private techniques. After careful analysis, it was discovered that different epsilon levels had a considerable impact on privacy and data utility metrics, particularly when it came to sensitive numerical components.

These findings have important applications for businesses that handle data security, privacy, and utility. The study provides useful guidance for enhancing comprehension and application of differential privacy techniques by highlighting the delicate balance that needs to be struck between these factors. To optimize privacy protection and data utility, this chapter emphasizes the significance of modifying privacy settings following organizational requirements and levels of data sensitivity.

4.1 Data Collection

4.1.1 Description of the Dataset

The "Adult Census Income" dataset was used for this study due to its invaluable resource for researching the outcomes of privacy preservation techniques because of its many characteristics, it's obtained from Kaggle. It includes a wide range of variables for analysis, including age, education level, marital status, capital gain, and more. Due to its size and complexity, the dataset is a good option for investigating the practical applications of differential privacy and other privacy-preserving techniques. Complete data preparation processes are necessary to guarantee the accuracy and quality of the data for analysis. Furthermore, the dataset's historical significance and comparability with more recent datasets are enhanced by its origins in the US Census database from 1994. This provides insights into the evolution of privacy concerns and data analysis over time.

4.1.2 Data Preprocessing

The dataset includes multiple attributes with missing values, such as "workclass," "occupation," and "native_country." The "workclass" attribute had 1,836 missing values, the "occupation" attribute had 1,843 missing values, and the "native_country" attribute had 583 missing values. These entries were specifically eliminated. The dataset decreased to 30,163 rows when the total number of removed records reached 2399. The validity and completeness of the dataset are guaranteed for further analysis through this careful data-cleaning procedure. It also included a range of numerical attributes that represent distinct facets of people's socioeconomic and demographic identity. Numerical values such as age, education level, weekly hours worked, capital gain, and capital loss are represented by these numerical values. By identifying the distribution and range of these variables among the people surveyed, each of these attributes offers important insights into the population included in the dataset. For example, the age attribute shows the age of each person, which may range from young adults to seniors. A person's education level, expressed in terms of education years, indicates how much of their formal education they have finished. Comparably, the number of hours worked per week measures the duration of time dedicated to work-related tasks. Financial aspects are represented by capital gain and loss, which show the gains or losses that people have made through investments or other financial activities. It is necessary to comprehend the nature of these numerical values to perform insightful analyses and draw conclusions from the dataset.

Further cleaning was done on the dataset to fix errors and inconsistencies after the entries with missing values were eliminated. This included resolving inconsistencies in numerical data, guaranteeing consistency in formatting, and standardizing categorical values. To ensure the dataset's accuracy and integrity and to provide a strong basis for insightful analysis and interpretation, these careful data preparation procedures are essential. Numerical attributes were also normalized for the evaluation phase to make sure they fall into a comparable range. The purpose of this normalization step was to evaluate various algorithms and metrics, thereby reducing measurement bias and improving the performance of analytical models. The accuracy and integrity of the dataset are ensured

by these meticulous data preparation steps, which also serve as a solid foundation for perceptive analysis and interpretation.

4.2 Overview of the Main Findings

The study's findings provide critical insights into the delicate balance between data utility and privacy protection within the framework of differentially private techniques. A thorough investigation revealed that differences in epsilon levels had a major effect on data utility and privacy measures, especially for sensitive numerical attributes. In particular, higher levels of privacy protection were consistently linked to lower epsilon values, which led to both a decrease in data utility and an increase in privacy gain. This emphasizes how difficult it is to choose epsilon values that best balance privacy and utility. A comprehensive evaluation of privacy and utility was made possible by the combination of metrics like Privacy Gain, Normalized Mutual Information (NMI), and Sum Squared Error (SSE). These measurements played a crucial role in calculating the trade-offs and directing the choice of epsilon values that offer the best possible balance. For example, the study showed that higher epsilon values improved the accuracy and usefulness of the data for analytical purposes, but they also decreased privacy protection.

To preserve privacy, we employed Laplace noise in our method to disturb the data. We successfully managed the trade-offs between privacy and data utility by carefully controlling the application of Laplace noise with the epsilon parameter. Through the application of curve fitting techniques, we were able to model the relationship between the hybrid metrics and epsilon values, thereby determining the epsilon value that best balances data utility and privacy protection.

The practical implications of these findings for organizations managing data security, privacy, and utility are noteworthy. Businesses will be better equipped to weigh the privacy vs utility trade-off when using the insights from this study to improve differential privacy technique understanding and implementation. The study offers practical advice for organizations attempting to manage the complexity of data management in a setting where privacy is becoming more and more important by illustrating the careful balance that must be struck between these factors.

The study also emphasizes how crucial it is to adjust privacy settings in accordance with particular organizational requirements and degrees of data sensitivity. The results indicate that, in order to maximize privacy protection and data utility, customized strategies should be used rather than a one-size-fits-all approach. This can contribute to a better-informed discussion on privacy-enhancing data usage in various sectors by informing policy-making, operational strategies, and the development of privacy-preserving technologies.

4.3 Applying Laplace Noise for Differential Privacy

Protecting individual privacy while preserving data utility requires applying differential privacy techniques, like adding Laplace noise, to sensitive attributes like age. We applied Laplace noise to the dataset and demonstrated how the amount of privacy protection applied can be controlled by adding Laplace noise to all values with different epsilon values such as age, education number, capital gain and loss, and hours per week. While masking individual-level details, this procedure guarantees that the perturbed data maintains statistical relevance and analytical utility. Also, we count the number of negative values in each perturbed column after it has been perturbed. this step is essential, to ensure that the privacy guarantees are maintained and to comprehend how a perturbation affects the distribution of data, this step is essential.

The dataset becomes more privacy-preserving when the attribute values are changed because the original values are obscured. However, there is a trade-off between privacy protection level and data utility. Greater privacy is possible, but the data's usefulness for some analyses may decrease as epsilon values increase due to the more obvious added noise. Given the particular use case and privacy requirements, this trade-off is crucial to privacy-preserving data analysis and needs to be carefully considered.

For example, between the original and perturbed age values for various epsilon values, the results in Table 2 offer a clear comparison. An evaluation of the effects of various privacy protection levels on the data is made possible by this comparison. This knowledge can help researchers and data analysts make decisions about how to balance privacy and utility by helping them understand how differential privacy techniques affect their data. In general, the use of Laplace noise to the age attribute highlights the larger

issues and concerns with privacy-preserving data analysis. It emphasizes how crucial it is to responsibly apply privacy-enhancing strategies to safeguard people's privacy while preserving the data's value for analysis and decision-making.

Table 3: Perturbed Ages of Individuals at Various Epsilon Values.

Age	0.1	0.2	0.3	0.5	1.0
82	89.262557	86.737559	79.746081	83.270263	83.643296
54	50.156174	57.090540	52.737222	54.775160	53.815080
41	41.795678	32.544603	37.242086	41.326009	39.907067
34	70.799462	36.344986	31.203465	34.236420	33.959483
38	64.701321	40.611345	39.923307	45.529068	38.341620
74	54.957611	83.561871	77.644861	72.971686	74.641017
68	68.970627	73.457163	74.042717	62.879799	68.976007
45	34.758331	45.692379	45.503620	43.980055	45.293184
38	31.638658	35.728152	39.792779	35.609037	39.122905
52	56.326689	55.454249	53.443465	52.089911	50.676903

Table 2 displays the original ages and perturbed ages of 10 individuals at different epsilon values (0.1, 0.2, 0.3, 0.5, and 1.0). Since the study used epsilon values ranging from 0.1 to 1.0 to fully investigate the effects, this table offers a sample representation. It shows that the perturbed ages generally decrease as the epsilon value increases, indicating higher privacy protection. This decrease is the outcome of more noise being added to the data to preserve sensitive data, which distorts the original values. Additionally, each individual's perturbed age varies for each of the various epsilon values shown in the data. The randomness of the noise created by the Laplace mechanism is reflected in this variability. It is important to note that although the perturbed ages show a noticeable change, the ages (from highest to lowest) maintain a stable order for everyone across a range of epsilon values. This shows that the relative relationships between ages remain

stable even after disturbance, emphasizing a compromise between the preservation of the data's overall structure and privacy protection. The data highlights the trade-off between privacy and data utility by showing how different privacy mechanisms, like the Laplace mechanism, affect specific data points.

4.4 Visualization

Understanding the effects of differential privacy strategies requires the use of visualization. The effects of applied Laplace noise on data distribution and utility are discussed in detail in the following sections.

The impact of differential privacy techniques—particularly the inclusion of Laplace noise—on the numerical attributes of the dataset are considered, producing informative visualizations. We can observe the trade-offs between privacy and data utility by comparing the original and perturbed values of each attribute across a range of epsilon values.

The perturbed values vary with varying levels of epsilon, as the epsilon decreases, and the line bar chart effectively illustrates this change. This graphic illustration makes it easier to comprehend how the distribution of data points within each attribute is impacted when Laplace noise is applied. The graph also makes it simple to compare different attributes, emphasizing those that are more concerned with protecting privacy. Higher sensitivity to privacy concerns may be indicated by attributes with a larger variance between the original and perturbed values across epsilon values. The impact of differential privacy on the dataset's numerical attributes is comprehensively outlined in these visualizations, which offer insightful information to researchers and practitioners who are trying to strike a balance between privacy and data utility in their analyses.

Figure 4 illustrates the perturbed ages at different epsilon values using a box plot. The visualization highlights that lower epsilon values are associated with more variation in the ages following the perturbation process, which suggests a higher degree of privacy protection. On the other hand, perturbed ages with higher epsilon values (e.g., 1.0) closely resemble the original values, and noise is the main source of variation. Increased variability in perturbed ages is the result of lower epsilon values (e.g., 0.1), which provide

stronger privacy protection at the risk of decreasing the data's analytical utility for some types of analyses.

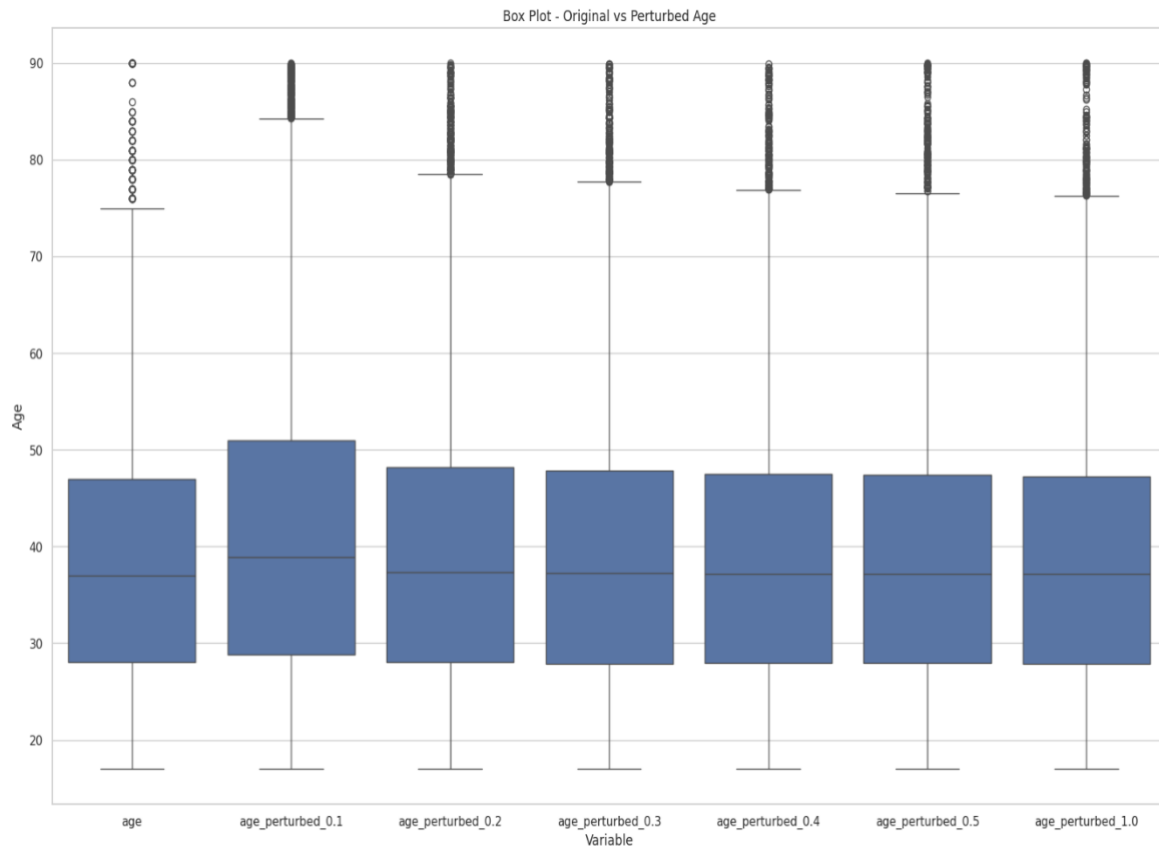


Figure 4: Effect of Epsilon Values on Age Perturbation: A Box Plot Analysis.

Interesting insights can be obtained from the box plot of the perturbed values of education numbers across various epsilon values as it's shown in Figure 5 In comparison to the original data, the distribution is noticeably skewed at epsilon 0.1, suggesting a significant distortion brought on by larger privacy protection. The distribution gradually becomes less skewed and approaches the original distribution's shape as epsilon increases. The box plot indicates a distribution that is slightly closer to the original by epsilon 1.0, indicating less distortion in the perturbed data. This finding implies a trade-off between data integrity and privacy protection. While lower epsilon values offer greater privacy protection, they introduce more distortion from the original data. Higher epsilon values, on the other hand, result in less distortion from the original data.

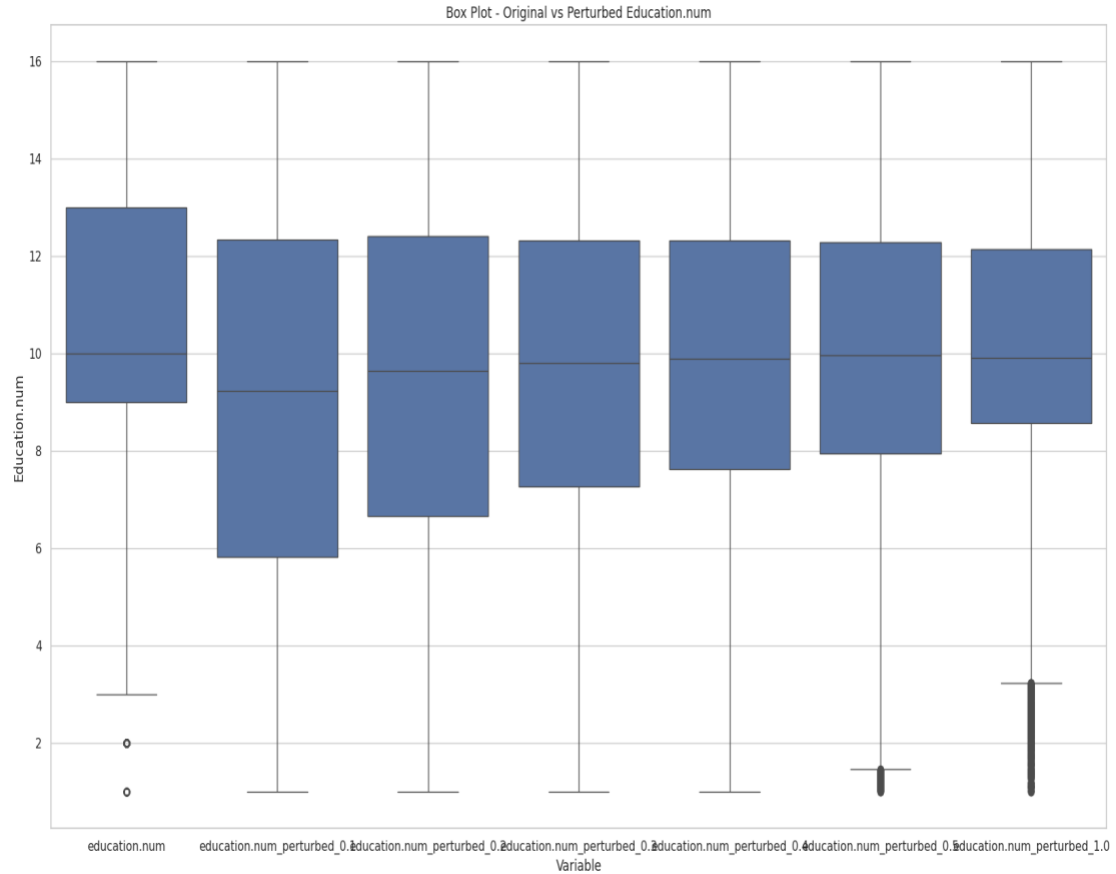


Figure 5: Effect of Epsilon Values on Education Numbers Perturbation: A Box Plot Analysis.

For every quantitative variable in the dataset, we create a histogram by comparing its original value side by side with the perturbed values at various epsilon levels for better comparison. We specify different bin widths based on the nature and range of the attribute. The variable capital loss employs a bin width of 500, which is probably chosen to align with the capital loss value range and guarantee sufficient coverage in the histogram. The bin width for capital gain is 1000 since the range of that column is from 0 to 99999. This expands the bin range and effectively captures the distribution of data. By considering the range and type of values found in the dataset, these bin width selections seek to produce histograms that accurately represent the data distribution for each variable.

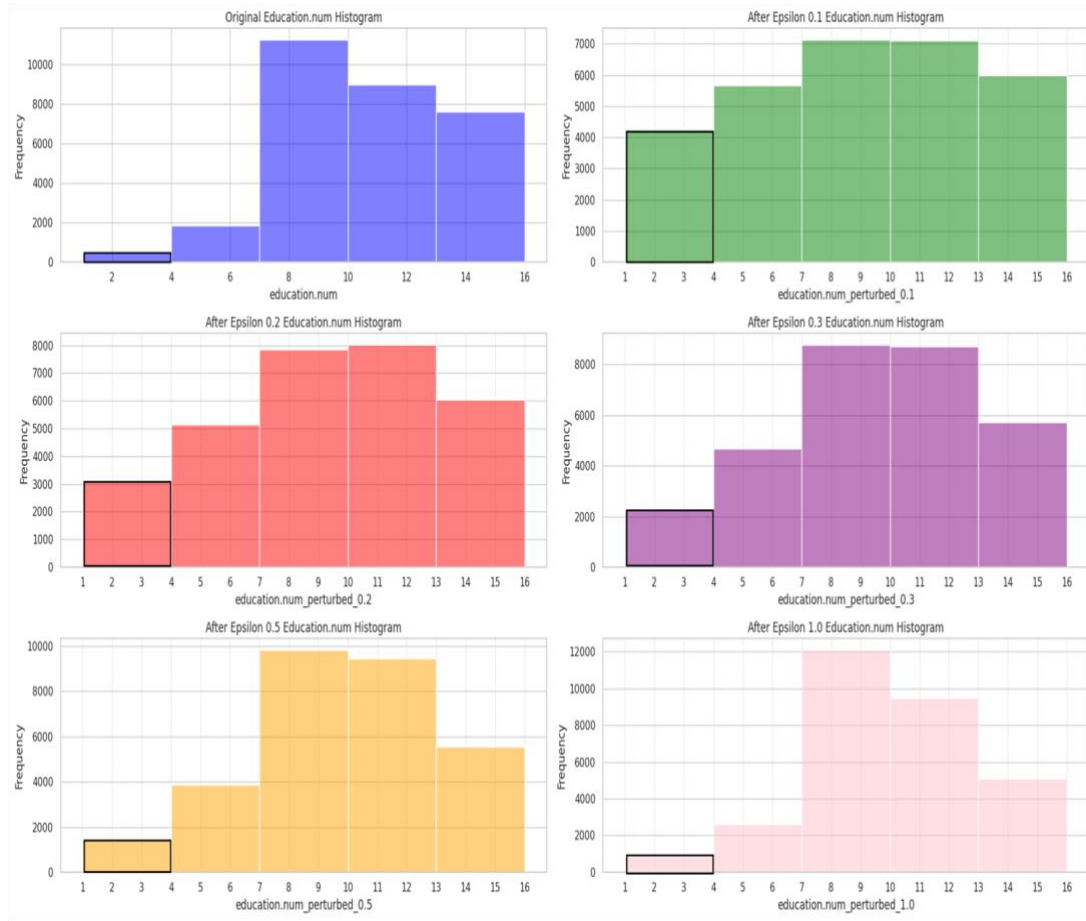


Figure 6: Comparison of Perturbed Education Numbers at Different Epsilon Values.

Figure 6 in various colors, indicating clear differences. There is a clear difference where there is a greater degree of privacy, particularly at an epsilon value of 0.1 where the bin heights vary significantly. To examine the comparison in greater detail, let's concentrate on the first bin. We note that in this case, the number of people with one to three years of education slightly exceeds 2000 following perturbation with epsilon 0.3, as opposed to the initial count of less than 2000. The count increases over 1000 at epsilon 0.5 and approaches the original count with epsilon 1.0, indicating less privacy protection. This analysis emphasizes the trade-off between data utility and privacy by highlighting the effects of varying epsilon values on the distribution of the perturbed data.

The trade-off between data utility and privacy is highlighted by the trends in the perturbed education numbers. More variability in the perturbed data is caused by a higher epsilon value, which may compromise the data's usefulness for some analyses but also reflects enhanced privacy protection. Lower epsilon values, on the other hand, maintain

data utility but provide less privacy protection because they produce perturbed data that closely resembles the original values.

The "Hours.per.week" variable's distribution is shown in Figure 7 both before and after Laplace noise is applied for various epsilon values. The original "Hours.per.week" data is represented, undisturbed, in the blue histogram. The distributions of "Hours.per.week" are shown in the following histograms, which were created by perturbing the data with epsilon values between 0.1 and 1.0. The corresponding epsilon value used for the perturbation is used to title each histogram.

The perturbed data distributions get closer to the original distribution as epsilon increases. The histograms exhibit a wider spread and a more significant deviation from the original data for lower epsilon values (e.g., 0.1, 0.2), which suggests higher levels of noise and, thus, greater privacy protection. The perturbed data distributions narrow and start to more closely resemble the original distribution as the epsilon value rises (for example, to 0.5, 0.6, and at last to 1.0). Higher epsilon values preserve the original data characteristics by reducing noise, but they provide less privacy protection. This trend illustrates the trade-off between privacy and data utility. The figure gives a clear picture of how different epsilon values affect the perturbed data's privacy-utility balance.

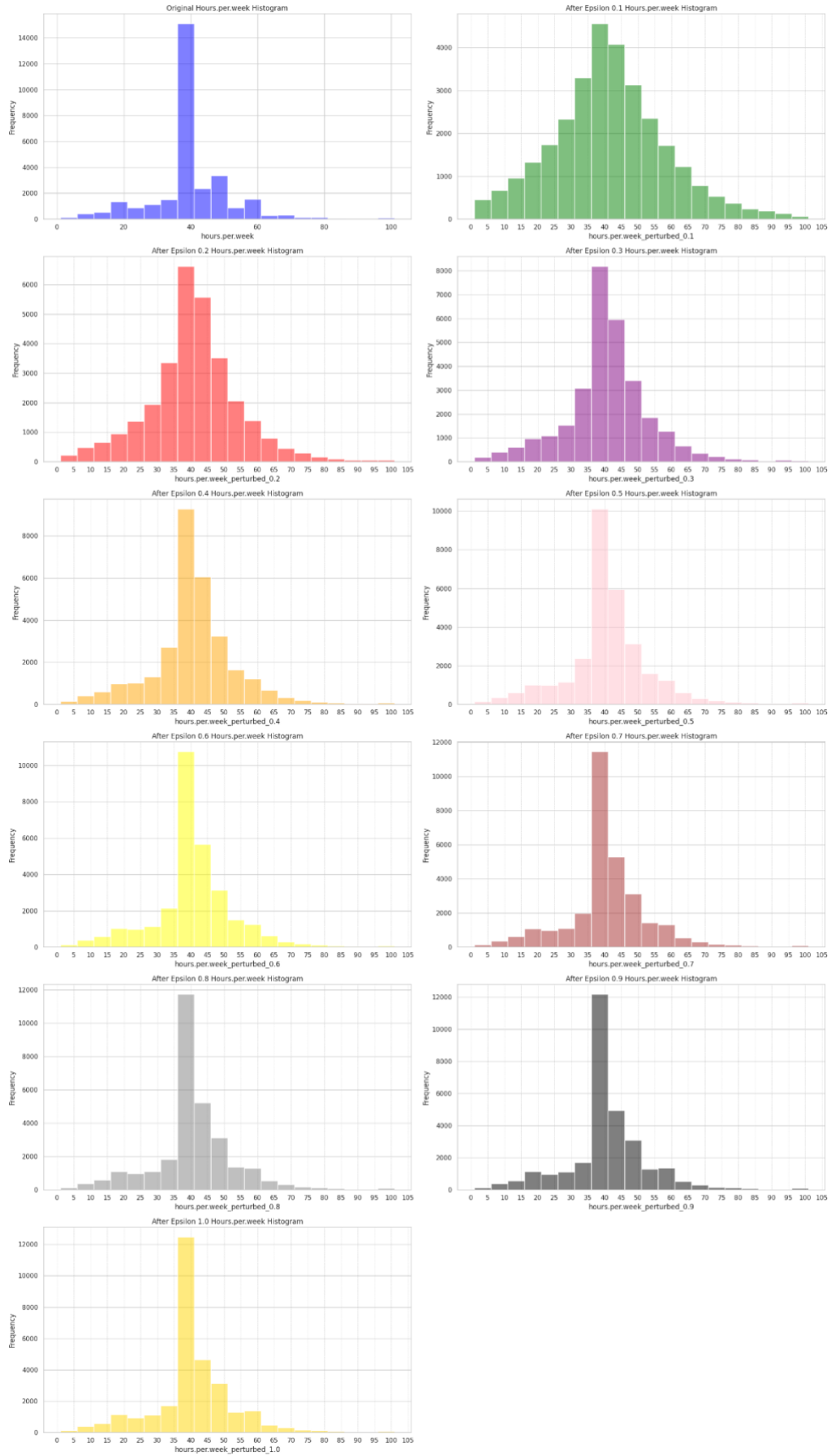


Figure 7: Comparison of Perturbed Hours Per Work at Different Epsilon.

Figure 8 compares the original bins with their perturbed counterparts using a different method. It is possible to compare the original and perturbed bins more directly with this method, which facilitates pinpointing the precise differences between them. Finding the epsilon value that achieves the best trade-off between privacy and utility is the goal of this comparison. The perturbed age at epsilon 0.5, for instance, shows that there was a change because of the perturbation because the bin at age 20 has a higher count in the perturbed data than in the original data. On the other hand, the perturbed data shows a lower count in the age 35 bin when compared to the original. However, when comparing the perturbed data to the original, the bins at ages 55, 65, 70, and 75 appear to be the same.

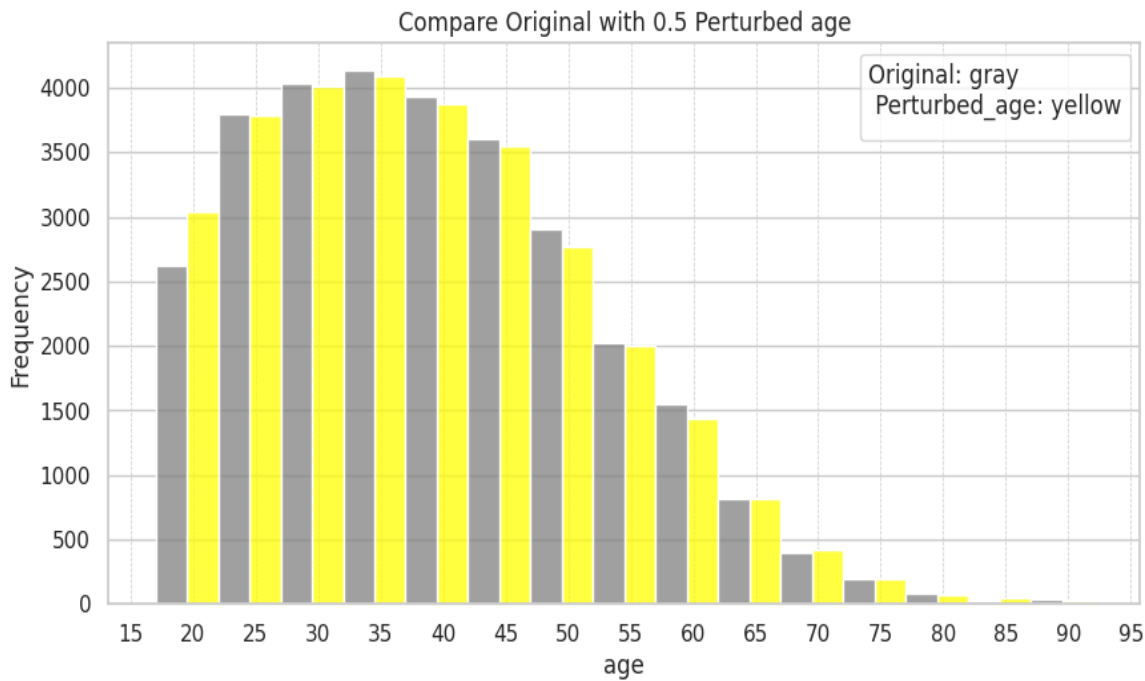


Figure 8: Comparison of Perturbed Age at 0.5 Epsilon Values.

Figure 9 analyzes another attribute, such as hours per week, to obtain more understanding of the impact of epsilon 0.5. Here, we find that the original and perturbed bins differ significantly, with a count difference of about 4000, at 40 hours per week. But the disturbed bins look a lot like the original bins at 10, 15, 30, and 70 hours a week. The context and requirements of the data analysis will determine whether epsilon 0.5 is a good trade-off. Stronger privacy protection is offered by larger epsilon values; however, they may also make the data less useful for certain types of analysis. Hence, depending on the

desired balance between privacy and utility in the particular use case, great thought should be given to the selection of epsilon. According to this, epsilon 0.5 preserves some usefulness in the data while offering a moderate level of privacy protection.

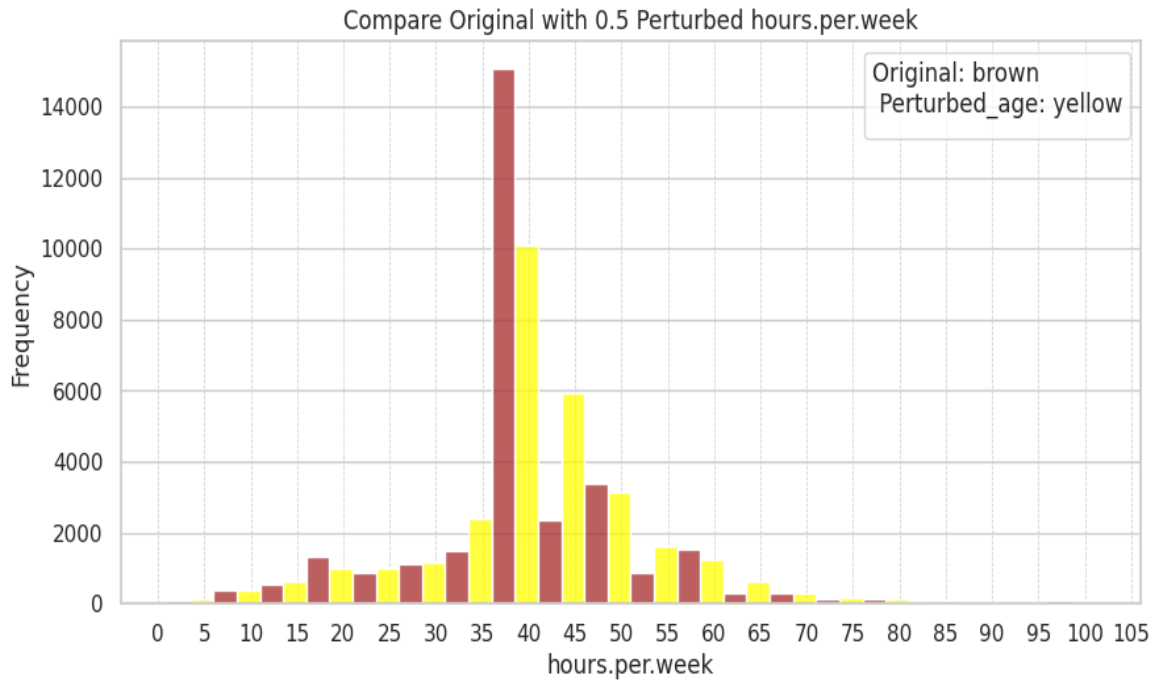


Figure 9: Comparison of Perturbed Hours Per Week at 0.5 Epsilon Values.

One interesting finding is that capital gains and losses appear to be more resilient to the effects of varying epsilon values. This resistance is seen in Figures 10 and 11, especially when it comes to capital gains, where the epsilon values are 0.1 and 0.3. There are very few observable differences between each bin in these figures and the original; in fact, changes are hardly noticeable at all. It's important to note that, to facilitate a clearer comparison, we removed from the visualization any original values that equal zero.

The nature of the data itself could be the root cause of this resistance. When it comes to capital gains and losses, a sizable percentage of the initial values are zero. For instance, only 2,712 of the dataset's 32,562 total values are nonzero; the remaining values are zeros. Due to the high prevalence of zero values, the data may be less variable and therefore more resistant to being perturbed by the addition of Laplace noise at various epsilon levels. For

this reason, even with different values of epsilon, the perturbed values for capital gain and loss are almost identical to the original values.

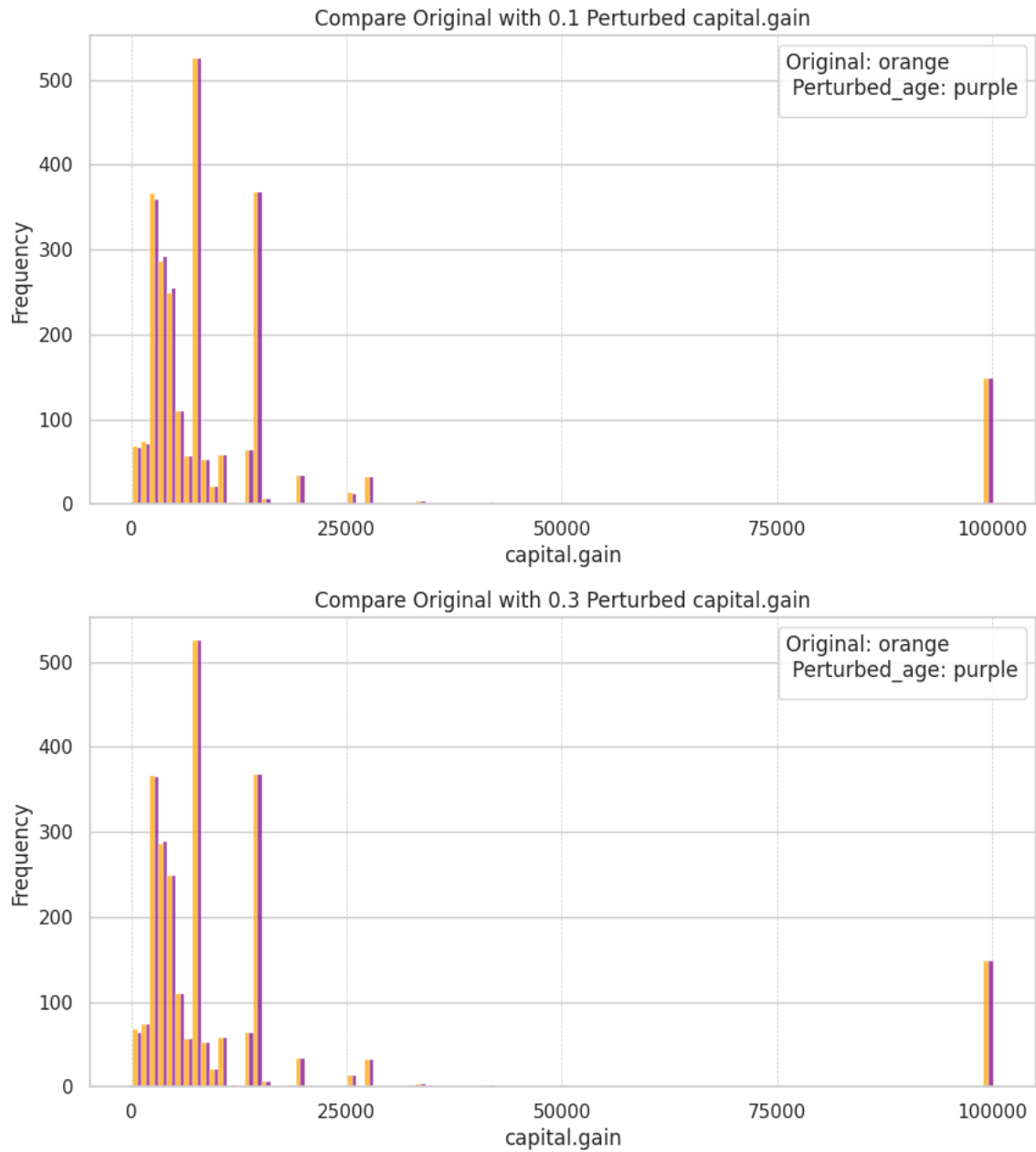


Figure 10: Comparison of Perturbed Capital Gain at Different Epsilon Value.

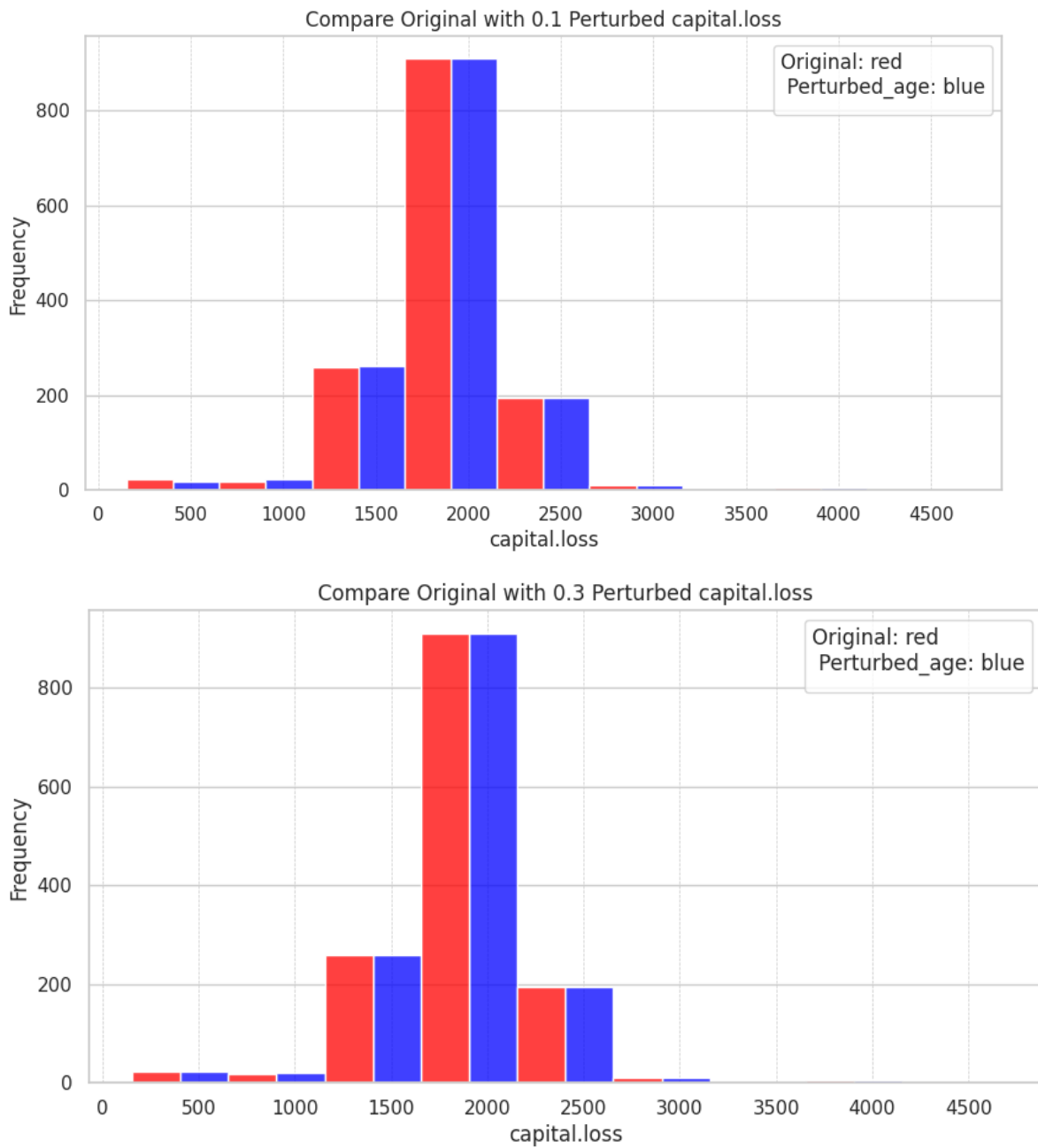


Figure 11: Comparison of Perturbed Capital Loss at Different Epsilon Value.

4.5 Evaluation Metrics

Evaluation metrics are critical to determining how well privacy-preserving strategies work. The following sub-sections look at the application of key metrics used in this study. We demonstrate the optimal value and find a compromise between data privacy and data utility by utilizing various evaluation metrics. These metrics offer quantifiable assessments of the trade-off between privacy and utility, such as privacy gain and NMI. We can identify the most appropriate epsilon value that minimizes NMI while maximizing the Privacy Gain by examining these metrics for different epsilon values and numerical attributes. This approach enables us to make well-informed decisions regarding the trade-off between privacy and utility, ensuring that the data maintains its analytical value while safeguarding the privacy of individuals.

Finding out how data perturbation techniques impact data privacy and utility requires an understanding of these evaluation metrics. Privacy Gain measures the enhancement in privacy protection that results from the perturbation process, taking into account the sensitivity of the data and the epsilon value chosen for perturbation. On the other hand, Normalized Mutual Information (NMI) measures how much information is retained in the perturbed data when compared to the original data. We can learn more about how different noise levels affect the privacy and usefulness of the data by methodically examining these metrics for a range of epsilon values and numerical characteristics. This analysis helps determine the optimal epsilon value for a given dataset and analytical task by balancing privacy and utility. These metrics allow comparisons of the efficacy of various perturbation strategies and help in the design of privacy-preserving mechanisms that maximize privacy gain and minimize the NMI while maintaining high levels of data utility. This section describes the techniques used to find the ideal epsilon value that strikes a compromise between utility and data privacy. We conducted two independent experiments: empirical and estimated experiments.

4.5.1 Applying Privacy Gain

The enhancement in privacy protection generated by the perturbation process is measured by privacy gain. We first apply Laplace noise to the data to calculate privacy gain, making sure the perturbed values stay within acceptable ranges. Next, both the original and perturbed data are scaled to the same value. The maximum absolute difference between the normalized original and perturbed data is used to calculate privacy gain. Figure 12 shows how privacy gain changes for age, education.num, and hours.per.week at different epsilon values.

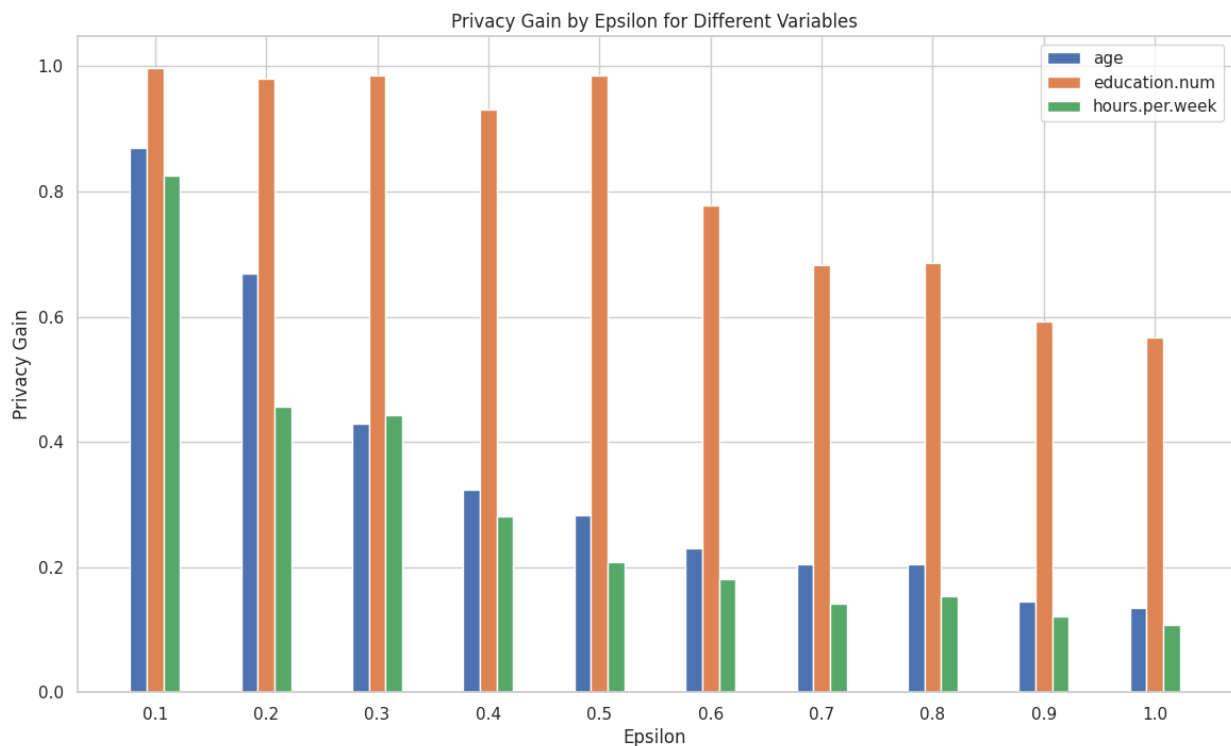


Figure 12: Privacy Gain by Epsilon for Different Variables.

The age variable exhibits strong privacy protection, as evidenced by the relatively high privacy gain of approximately 0.87 at the lowest epsilon value of 0.1. The privacy gain gradually declines as epsilon rises, suggesting that the disturbance has less of an effect on privacy. The privacy gain decreases to about 0.13 by the time epsilon reaches 1.0, indicating less privacy protection. According to this trend, there are significant privacy gains for the age variable at smaller epsilon values; however, the benefit decreases with

increasing epsilon. This implies that when choosing epsilon for age, the trade-off between privacy and utility needs to be carefully considered.

Between all epsilon values, the education.num variable exhibits the largest privacy gain, beginning at almost 1.0 for epsilon 0.1. This suggests that, at low epsilon values, the perturbation offers the education.num variable strong privacy protection. In comparison to the other variables, the privacy gain stays relatively high even as epsilon increases. For example, the privacy gain remains approximately 0.98 at epsilon 0.5. It decreases to roughly 0.57 by epsilon 1.0, indicating that even at higher epsilon values, the variable maintains a sizable degree of privacy protection. The high privacy gain for education.num indicates that privacy-preserving techniques are more beneficial for this variable because it is more sensitive to perturbations. Nevertheless, significant amounts of noise introduction may also have an impact on the data utility.

At lower epsilon values, the hours.per.week privacy gain is likewise high; it begins at about 0.83 for epsilon 0.1. As epsilon rises, the privacy gain falls off similarly to the other variables. The privacy gain is approximately 0.21 at epsilon 0.5 and decreases to approximately 0.11 at epsilon 1.0. This shows that hours.per.week starts off with strong privacy protection, but it decreases significantly faster than education.num as epsilon values rise. The drop in privacy gain for hours.per.week emphasizes the necessity of striking a balance between privacy and usefulness, especially for variables for which usefulness is essential to analysis. It may be necessary to use lower epsilon values to protect privacy without significantly sacrificing data utility.

The results of privacy gain for the education.num are shown in Table 3 across various epsilon values. Through the perturbation process, privacy protection has improved, as measured by privacy gain. The privacy gain typically declines as the epsilon value rises, indicating a decrease in privacy protection.

Higher privacy gains are associated with lower epsilon values, suggesting stronger privacy protection, according to the analysis of the privacy gain chart. Across all epsilon values, education.num exhibits the highest privacy gain among the variables, indicating that it is the one that gains the most privacy protection from the Laplace noise perturbation. On the other hand, although age and hours.per.week show significant increases in privacy

at lower epsilon values, the rate at which their privacy protection decreases with increasing epsilon is higher. This analysis facilitates comprehension of the responses of various variables to privacy-preserving strategies and aids in determining the ideal trade-off between privacy and utility.

Table 4: Privacy Gain Results for education.num.

Epsilon	Privacy Gain
0.1	0.998005
0.2	0.979817
0.3	0.985801
0.4	0.931214
0.5	0.984668
0.6	0.777150
0.7	0.681798
0.8	0.686826
0.9	0.592840
1.0	0.566785

4.5.2 Applying NMI

The amount of information retained in the perturbed data relative to the original data is measured by Normalized Mutual Information (NMI). First, we use min-max normalization to normalize the original and perturbed data to calculate NMI. Next, the discretized original and perturbed data are divided into bins using the normalized data, and the NMI is calculated. Greater NMI values represent enhanced information retention of the original data following disturbance.

The NMI values for the variables education.num, hours.per.week, and age are shown across various epsilon levels in Table 4. This table shows in detail how information retention gets better as the epsilon value rises. For example, the NMI for education.num is very low at 0.011670 at epsilon 0.1, suggesting that the noise added causes most of the information to be lost. On the other hand, education.num's NMI of 0.373906 at epsilon 1.0 indicates a notable improvement in information retention. Similarly, when less noise is added, the hours.per.week variable demonstrates better retention, progressing from

0.160823 at epsilon 0.1 to 0.653625 at epsilon 1.0. As the perturbation decreases, the age variable, which has the highest NMI values, increases from 0.162595 at epsilon 0.1 to 0.763038 at epsilon 1.0, demonstrating its resilience in retaining information.

Table 5: NMI Results for Different Epsilon Values.

Epsilon	Variable	NMI
0.1	education.num	0.011670
0.1	hours.per.week	0.160823
0.1	age	0.162595
0.2	education.num	0.041800
0.2	hours.per.week	0.347454
0.2	age	0.357541
0.3	education.num	0.081916
0.3	hours.per.week	0.460653
0.3	age	0.478549
0.4	education.num	0.127998
0.4	hours.per.week	0.529508
0.4	age	0.565493
0.5	education.num	0.174724
0.5	hours.per.week	0.577069
0.5	age	0.621282
0.6	education.num	0.219034
0.6	hours.per.week	0.604466
0.6	age	0.664778
0.7	education.num	0.264395
0.7	hours.per.week	0.622609
0.7	age	0.693642
0.8	education.num	0.299529
0.8	hours.per.week	0.636698
0.8	age	0.724776
0.9	education.num	0.337182
0.9	hours.per.week	0.647630
0.9	age	0.750452
1.0	education.num	0.373906
1.0	hours.per.week	0.653625
1.0	age	0.763038

Figure 13 shows how the variables education.num, hours.per.week, and age change as NMI varies across epsilon values. The NMI for the education.num variable is extremely low at epsilon 0.1, or roughly 0.01; this suggests that information retention is poor at this

degree of disturbance. On the other hand, the NMI rises dramatically with increasing epsilon, reaching roughly 0.37 at epsilon 1.0. This pattern implies that better information persistence of the original data for education.num is associated with higher epsilon values (less noise).

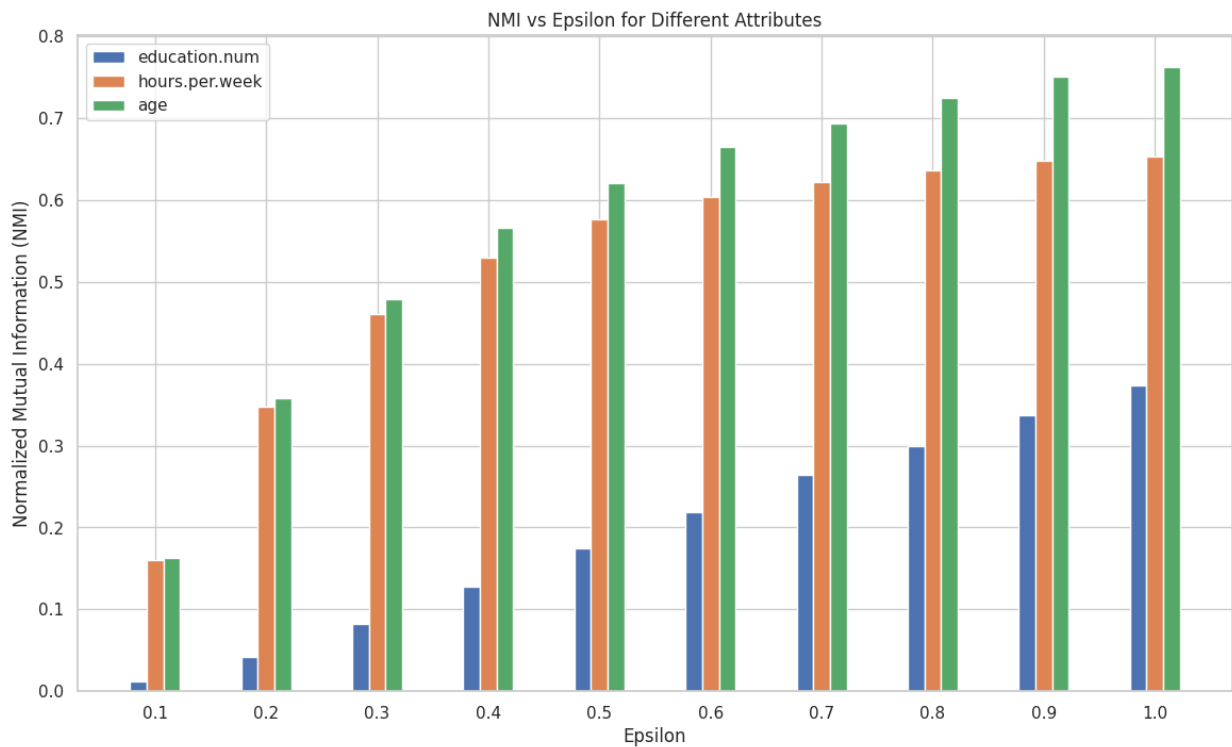


Figure 13: NMI by Epsilon for Different Variables.

In contrast to education.num, the hours.per.week variable displays a similar pattern, but with higher NMI values at each epsilon level. The NMI is about 0.16 at epsilon 0.1. The NMI rises gradually with increasing epsilon, reaching roughly 0.65 at epsilon 1.0. This suggests that at all perturbation levels, hours.per.week retains more information than education.num.

Among the three variables for the age variable, the NMI is the highest across all epsilon values. The NMI increases to roughly 0.76 at epsilon 1.0 from a starting point of roughly 0.16 at epsilon 0.1. This suggests that the age variable retains more information even at higher noise levels and is less perturbed than the other variables.

The NMI chart analysis shows that the NMI values for all variables increase as epsilon increases, indicating less noise. This pattern implies that greater epsilon values result in greater information retention from the original data. Age is the variable that constantly exhibits the highest NMI values, meaning it retains the most information following disturbance. Education.num has the lowest NMI values across all epsilon levels, followed by hours.per.week and education.num, which have slightly lower values. To balance the trade-off between privacy and data utility and guarantee that information retention satisfies the analysis's requirements, this analysis emphasizes the significance of choosing suitable epsilon values.

4.5.3 Empirical Exploration Results

This sub-section explores the findings derived from empirical examination of the data in various privacy configurations, emphasizing the efficacy of hybrid metrics.

4.5.3.1 Hybrid Metric for Privacy Gain and SSE

In this work, we examined a hybrid metric that balances privacy gain and utility loss to determine the ideal epsilon (ϵ) values for various attributes, including education.num, hours.per.week, and age. To comprehend the trade-offs between these two factors, this involved varying privacy and utility weight combinations, ranging from privacy weight = 0 to privacy weight = 1.0.

Laplace noise was first applied to the original data to disturb it and make sure the disturbed values stayed within the range of the original data. To simulate data anonymization and safeguard privacy without significantly reducing the usefulness of the data, this step was essential. Following that, the perturbed and original data were both normalized to a similar scale so that privacy gain and utility loss could be computed more easily.

Privacy Gain is calculated to quantify the degree of data alteration and thus the degree of privacy protection attained, the maximum absolute difference between the normalized original and perturbed data. The Sum of Squared Errors (SSE) between the normalized frequencies of the original and perturbed data was used to calculate the utility

loss. This metric was selected because it accurately measures the utility loss brought on by data perturbation. We also calculated the average SSE and root sum of squares to gain an additional understanding of the impact of the data perturbation.

The bin widths were manually specified for each attribute: age was assigned 22 bins, education.num and hours.per.week were assigned 9 bins each. The empirical specification ensured that the frequency counts were comparable and meaningful by basing it on the type and distribution of data in these columns.

Next, we calculated a hybrid metric using different weights for privacy and utility, which is a weighted sum of privacy gain and utility loss ($1 - \text{SSE}$). We were able to investigate various trade-offs between privacy and usefulness because of this. We performed constrained optimization to make sure epsilon stayed within the $[0.1, 1.0]$ range and visualized the result of the hybrid metric in a line chart to determine the optimal epsilon value.

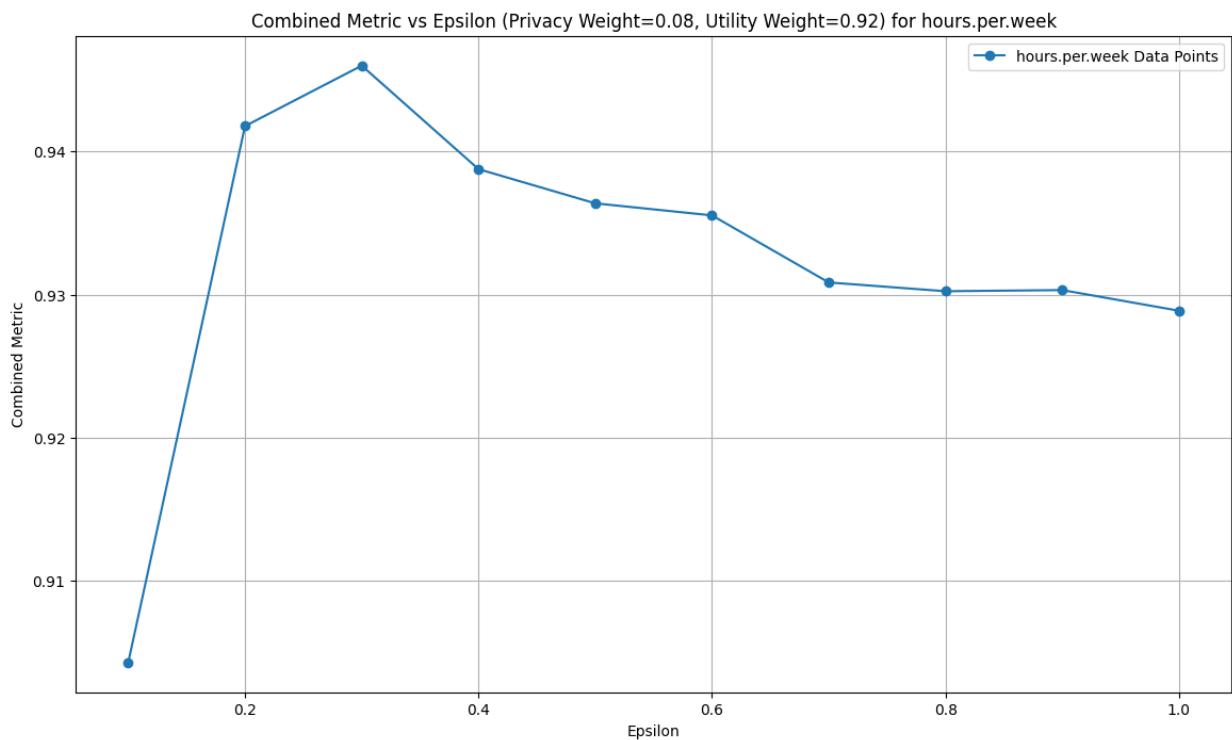


Figure 14: Hybrid Metric for PG and SSE for hours.per.work.

In the beginning, we looked at each attribute's behavior separately to find the best weight combinations to strike a balance between privacy and utility. Figure 14 shows the hybrid metric with different epsilons for the attribute "hours.per.week" with a privacy weight of 0.08 and a utility weight of 0.92. This particular weight combination indicates that data utility is prioritized over privacy. The hybrid metric in this case peaks at an epsilon value of about 0.3, indicating that this epsilon value offers the best compromise between preserving data utility and guaranteeing privacy protection. Beyond this point, the hybrid metric decreases, suggesting that while increasing epsilon further improves privacy, it also begins to significantly reduce data utility.

Similar analyses are shown in Figure 15, although the privacy and utility weight combinations vary. These variations serve as an example of how changing the weight priorities can affect the optimal epsilon value even for the same attribute. In the chart where the utility weight is set at 0.85 and the privacy weight at 0.15. The optimal balance point is sensitive to the specified weights, as further evidenced by the fact that the hybrid metric's peak occurs at a different epsilon value which is 0.2. The hybrid metric again peaks at the same epsilon value of 0.2, with privacy weight of 0.20 and utility weight of 0.80. However, with a utility weight of 0.90 and a privacy weight of 0.10, the hybrid metric peaks at 0.3 same as Figure 14. This change implies that the ideal epsilon value required to obtain the best balance can be changed by making small adjustments to the relative importance of privacy versus utility.

These findings demonstrate that the ideal epsilon value is not constant but rather depends on the desired trade-off between privacy and utility. The observed variations between the charts indicate that practitioners should carefully choose and modify these weights according to their unique needs and the circumstances surrounding their data analysis assignments. Because of its flexibility, a customized strategy for striking a balance between privacy and utility can be implemented, guaranteeing that the chosen epsilon value will satisfy the requirements of the application while maintaining the data's usefulness and safeguarding the privacy of everyone.

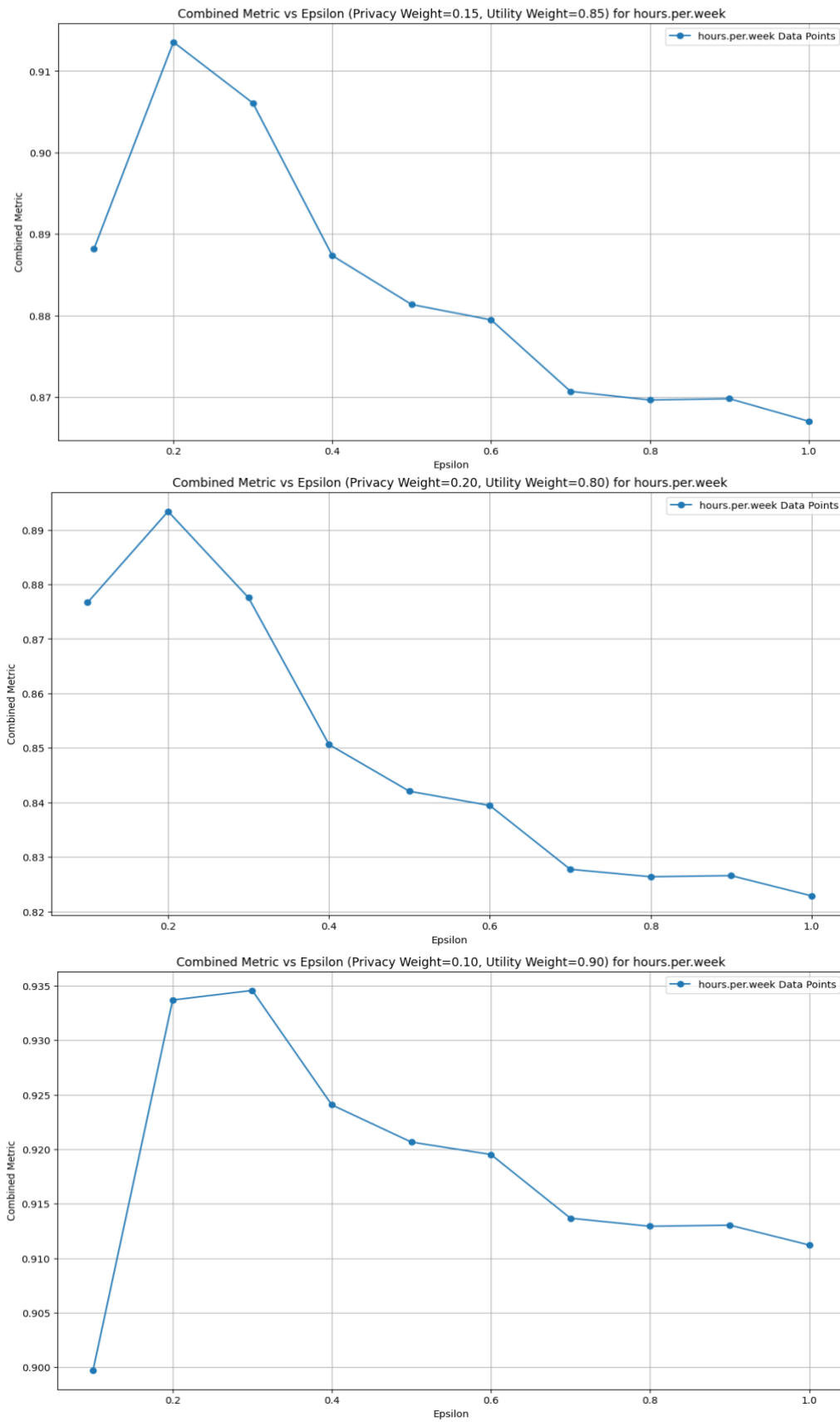


Figure 15: Hybrid Metric for PG and SSE for hours.per.work in Different Combinations of Weights.

To find the optimal epsilon values for each attribute, we used Laplace noise to perturb the original data over a range of epsilon values while maintaining the perturbed data's range within the original dataset. We determined the privacy gain for each perturbed dataset. This gives an indication of the extent to which the data has been modified for privacy protection. Furthermore, we computed the sum of squared errors (SSE) between the normalized frequencies of the perturbed and original data to evaluate the utility loss. The SSE aids in calculating how much a perturbation affects the utility of data.

By calculating a hybrid metric for each combination of privacy and utility weights, we were able to balance the gain in privacy and the loss in utility. Finding the highest point in the combined metric line charts when merging privacy gain and accuracy allowed us to calculate the ideal epsilon values for each attribute. The lowest point in the hybrid metric line charts was used to calculate the ideal epsilon when combining NMI and SSE.

The findings of the analysis indicate that when the utility weight declines and the privacy weight rises, the hybrid metrics' peaks typically move to lower epsilon values. While there are some exceptions based on sensitivity, this pattern is generally consistent across all attributes. While the age and education.num attributes show more consistent trends, the hours.per.week attribute displays multiple peaks, suggesting a more complex relationship with epsilon.

Table 6 summarize the best epsilon values for every characteristic across various privacy and utility weight combinations. The best epsilon value for each attribute and weight combination is shown in this table, enabling the best possible trade-off between privacy and utility. The best epsilon is represented by the highest point in the line chart when privacy gain and accuracy are combined. This in-depth knowledge is essential for modifying epsilon values to attain the appropriate balance between privacy and usefulness, guaranteeing successful data anonymization and subsequent analysis.

Table 6: Best Epsilon Values for Each Attribute.

	Column	Privacy Weight	Utility Weight	Best Epsilon
0	education.num	0.00	1.00	1.0
1	hours.per.week	0.00	1.00	0.7
2	age	0.00	1.00	1.0
3	education.num	0.01	0.99	1.0
4	hours.per.week	0.01	0.99	0.7
5	age	0.01	0.99	0.1
6	education.num	0.02	0.98	1.0
7	hours.per.week	0.02	0.98	0.7
8	age	0.02	0.98	0.1
9	education.num	0.03	0.97	1.0
10	hours.per.week	0.03	0.97	0.4
11	age	0.03	0.97	0.1
12	education.num	0.05	0.95	0.8
13	hours.per.week	0.05	0.95	0.4
14	age	0.05	0.95	0.1
15	education.num	0.08	0.92	0.8
16	hours.per.week	0.08	0.92	0.3
17	age	0.08	0.92	0.1
18	education.num	0.10	0.90	0.5
19	hours.per.week	0.10	0.90	0.3
20	age	0.10	0.90	0.1
21	education.num	0.15	0.85	0.5
22	hours.per.week	0.15	0.85	0.2
23	age	0.15	0.85	0.1
24	education.num	0.20	0.80	0.5
25	hours.per.week	0.20	0.80	0.1
26	age	0.20	0.80	0.1
27	education.num	0.30	0.70	0.5
28	hours.per.week	0.30	0.70	0.1
29	age	0.30	0.70	0.1
30	education.num	0.40	0.60	0.1
31	hours.per.week	0.40	0.60	0.1
32	age	0.40	0.60	0.1
33	education.num	0.50	0.50	0.1
34	hours.per.week	0.50	0.50	0.1
35	age	0.50	0.50	0.1
36	education.num	0.60	0.40	0.1
37	hours.per.week	0.60	0.40	0.1
38	age	0.60	0.40	0.1
39	education.num	0.70	0.30	0.1
40	hours.per.week	0.70	0.30	0.1
41	age	0.70	0.30	0.1
42	education.num	0.80	0.20	0.1
43	hours.per.week	0.80	0.20	0.1
44	age	0.80	0.20	0.1
45	education.num	0.90	0.10	0.1
46	hours.per.week	0.90	0.10	0.1
47	age	0.90	0.10	0.1

4.5.3.2 Hybrid Metric for NMI and SSE

This study also investigates the correlation between normalized mutual information (NMI) and the sum of squared errors (SSE) for the same attributes, at varying epsilon values. Understanding the trade-offs between privacy and utility in data anonymization requires an understanding of these metrics. To ensure that the perturbed values stayed within the original data range, the approach involved perturbing the original data with Laplace noise as we did with the privacy gain. The original and perturbed data were then normalized to make the computation of NMI and SSE simpler.

The mutual information score was computed after the normalized original and perturbed data were discretized into bins to calculate NMI. The metric measures the degree of resemblance between the original and perturbed data distributions; higher values correspond to decreased data utility retention. Lower values indicate less utility loss. In contrast, SSE measures the deviation between the original and perturbed data frequencies.

To perform the analysis, different combinations of privacy and utility weights were tested, ranging from privacy weight = 0 to privacy weight = 1.0. A comprehensive measure of the trade-off between NMI and utility loss was obtained by computing a hybrid metric, which was calculated as a weighted sum of NMI and SSE. By fitting a quadratic function to the hybrid metric values and identifying the peaks of the fitted curves, the ideal epsilon values were found as we did before. By guaranteeing that the epsilon values fell between [0.1, 1.0], this technique successfully balanced data utility and privacy protection.

An example of the education.num attribute's hybrid metric versus epsilon is shown in Figure 16, where the utility weight is 0.85 and the privacy weight is 0.15. This hybrid metric combines the concepts of Sum of Squared Errors (SSE) and Normalized Mutual Information (NMI). A lower total metric value in this case refers to a better trade-off between privacy and usefulness. The U-shaped pattern on the curve indicates the ideal epsilon value for this set of weights at its lowest point. At epsilon = 0.1, the curve begins with a relatively high hybrid metric. The hybrid metric falls with an increase in epsilon and reaches its minimum at epsilon = 0.3 to 0.4. The hybrid metric starts to rise again after this point.

The ideal balance between NMI and SSE is represented by the lowest point on the curve, which happens at $\epsilon = 0.4$. The fact that the hybrid metric value has reached its minimum at this point suggests that this epsilon value offers the best compromise between reducing noise-induced distortion (measured by SSE) and maintaining the original data's structure (measured by NMI). A lower hybrid metric value indicates that the data's structure is not unduly distorted by the noise added to it, all the while maintaining a suitable level of privacy. The selected epsilon value of 0.4 guarantees a reasonable degree of privacy protection while ensuring that the data is still useful for analysis.

Beyond $\epsilon = 0.4$, the hybrid metric value increases, suggesting that as epsilon increases, less noise is added to the data, resulting in less privacy protection and a higher SSE. As a result, the trade-off is less advantageous. On the other hand, excessive noise is produced by extremely low epsilon values (such as 0.1), which greatly distorts the data and raises the hybrid metric. In summary, with the given privacy and utility weights, 0.4 is the ideal epsilon value for the education.num attribute. The best compromise between preserving data utility and guaranteeing privacy is offered by this value.

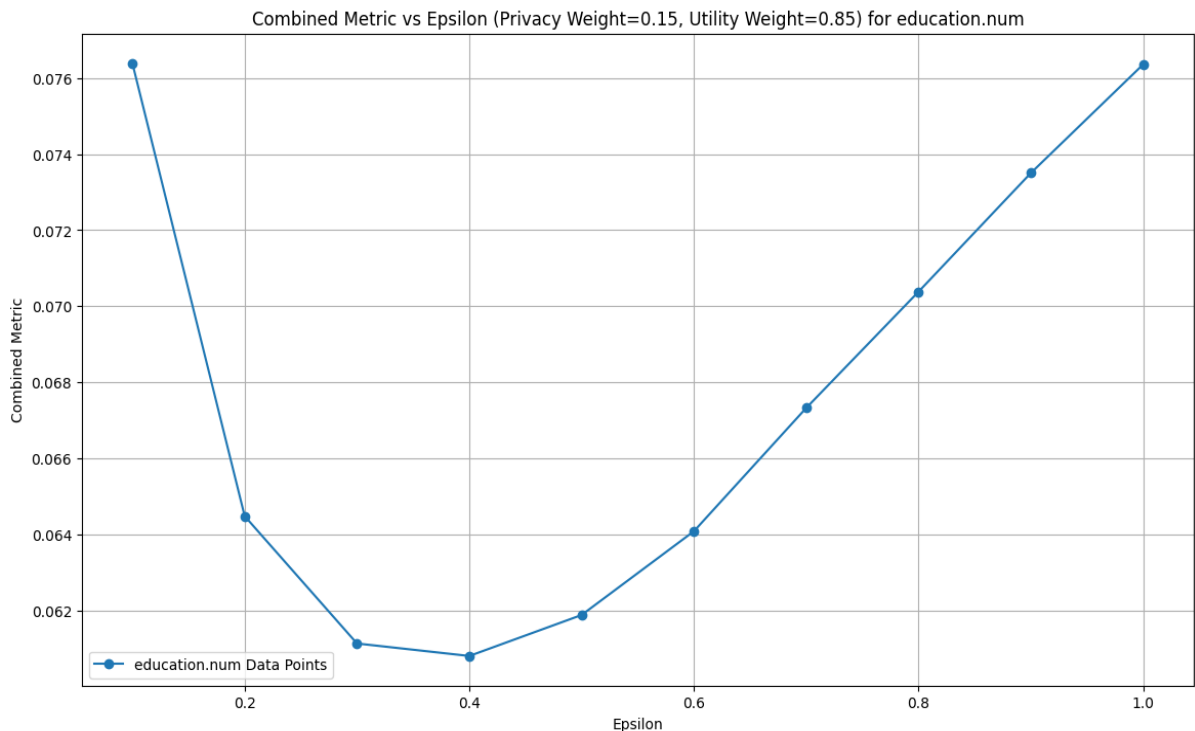


Figure 16: Optimal Epsilon Value Determination for education.num Using Combined Metric.

The hybrid metric for the education.num attribute versus epsilon for various combinations of privacy and utility weights are shown in Figure 17. The combinations shown in the charts are those in which the utility and privacy weights are, respectively, 0.20 and 0.80 and 0.30 and 0.70.

In the figure where the utility weight is 0.70 and the privacy weight is 0.30, the total metric increases steadily as epsilon rises. This suggests that lower epsilon values work better in this combination; the ideal epsilon is found at the lowest point of the combined metric curve, which is 0.2. To maintain an ideal balance, a smaller epsilon is required because this combination stresses privacy a little more. The hybrid metric in the second chart in Figure 17 also shows a U-shaped curve, with a privacy weight of 0.20 and a utility weight of 0.80. This curve's lowest point is located at $\epsilon = 0.3$. This suggests that the optimal epsilon stays close to the same value even when the emphasis on privacy is slightly increased in comparison to the first combination (0.15 privacy weight and 0.85 utility weight).

These graphs show that, even for the same attribute, the ideal epsilon value can differ considerably depending on the weight combination selected. The ideal epsilon value typically moves towards lower values as the privacy weight rises and the utility weight falls. This change reflects a stronger focus on adding noise to preserve privacy, which is consistent with the combined metric's higher priority for privacy. On the other hand, larger utility weights balance towards better data utility preservation by producing higher optimal epsilon values. In result, different combinations of privacy and utility weights result in different values for the optimal epsilon for the education.num attribute. When privacy is the top priority, lower epsilon values are better; conversely, when utility is the focus, higher epsilon values are better. These results emphasize how crucial it is to carefully choose epsilon depending on the privacy and utility balance needed for the analysis.

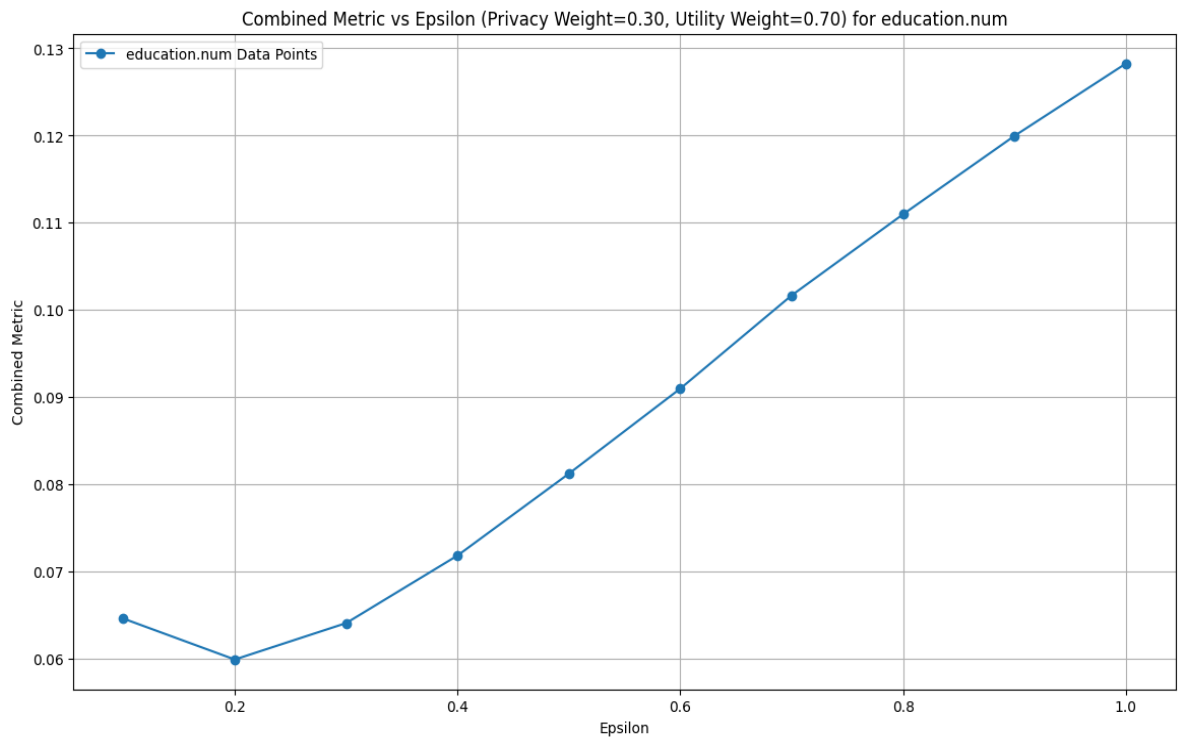
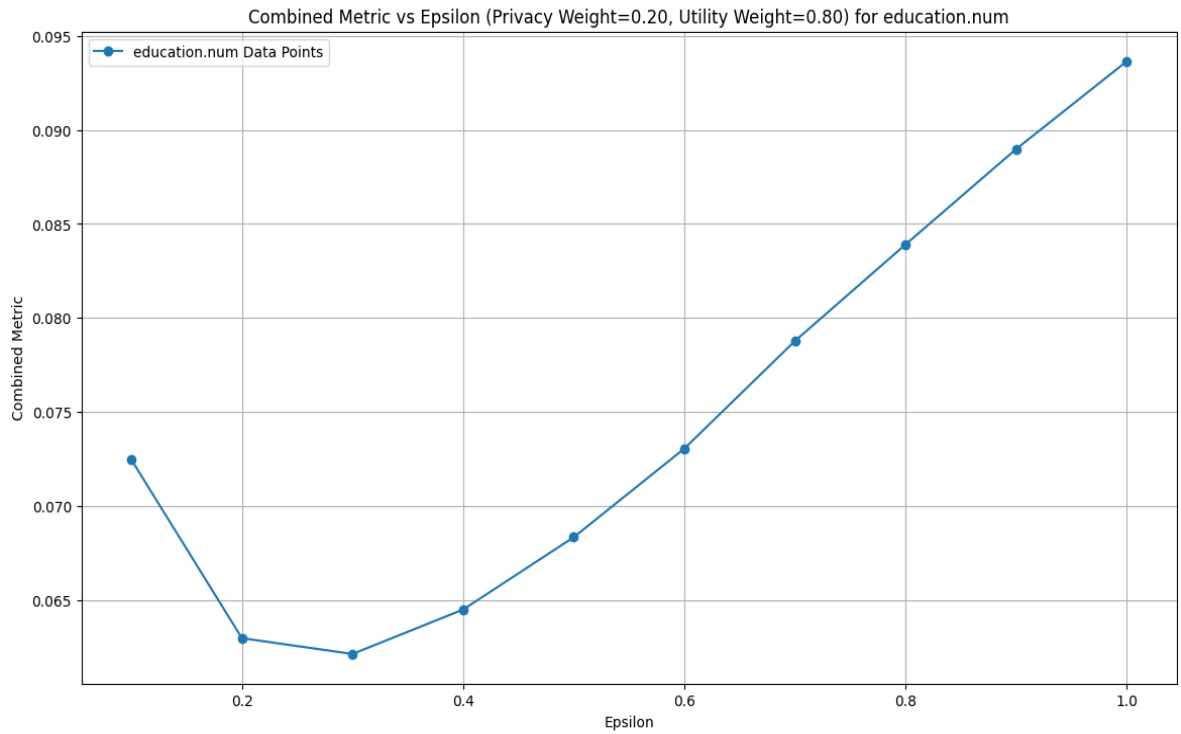


Figure 17: Optimal Epsilon Value Determination for education.num Using Combined Metric for Different Combinations of Weights.

The ideal epsilon values for each attribute (education.num, hours.per.week, and age) are displayed in Table 7 for various combinations of privacy and utility weights. To minimize the combined metric, which balances NMI and SSE, the ideal epsilon values were determined using the hybrid metric.

For the education.num attribute, the ideal epsilon value is 1, indicating little noise addition, when the utility weight is high (e.g., 1.0). The optimal epsilon falls as the privacy weight rises, indicating that more noise is added to preserve privacy. For instance, the ideal epsilon is 0.2 when the privacy weight is 0.30. When the privacy weight is 0.4 or more, the smallest optimal epsilon (0.1) is seen.

For the hours.per.week attribute, a high utility weight (1.0) results in an optimal epsilon value of 0.6. As the privacy weight increases, the optimal epsilon value decreases. For instance, with a privacy weight of 0.10, the optimal epsilon is 0.3. The lowest epsilon value (0.1) is seen for privacy weights of 0.20 and higher.

Similarly, for the age attribute, a high utility weight (1.0) corresponds to an optimal epsilon value of 0.6. As the privacy weight increases, the optimal epsilon value decreases significantly. For example, with a privacy weight of 0.01, the optimal epsilon is 0.1. The lowest epsilon value (0.1) appears consistently for privacy weights of 0.01 and higher.

In summary, the analysis shows that depending on how privacy and utility are balanced, different epsilon values are ideal. Higher epsilon values, which indicate less noise addition and higher data utility, are typically the result of higher utility weights. On the other hand, lower epsilon values correspond to higher privacy weights, suggesting that more noise should be added to improve privacy protection. This pattern emphasizes how crucial it is to carefully choose epsilon depending on the particular privacy-to-utility ratio that the analysis calls for. This table offers a helpful guide for selecting the right epsilon value in practical applications so as to attain the intended trade-off between privacy protection and data utility.

Table 7: Optimal Epsilon Values for Different Privacy and Utility Weights Across Attributes.

	Column	Privacy Weight	Utility Weight	Best Epsilon
0	education.num	0	1	1
1	hours.per.week	0	1	0.6
2	age	0	1	0.6
3	education.num	0.01	0.99	1
4	hours.per.week	0.01	0.99	0.5
5	age	0.01	0.99	0.1
6	education.num	0.02	0.98	1
7	hours.per.week	0.02	0.98	0.5
8	age	0.02	0.98	0.1
9	education.num	0.03	0.97	1
10	hours.per.week	0.03	0.97	0.4
11	age	0.03	0.97	0.1
12	education.num	0.05	0.95	1
13	hours.per.week	0.05	0.95	0.4
14	age	0.05	0.95	0.1
15	education.num	0.08	0.92	0.7
16	hours.per.week	0.08	0.92	0.3
17	age	0.08	0.92	0.1
18	education.num	0.1	0.9	0.6
19	hours.per.week	0.1	0.9	0.3
20	age	0.1	0.9	0.1
21	education.num	0.15	0.85	0.4
22	hours.per.week	0.15	0.85	0.2
23	age	0.15	0.85	0.1
24	education.num	0.2	0.8	0.3
25	hours.per.week	0.2	0.8	0.1
26	age	0.2	0.8	0.1
27	education.num	0.3	0.7	0.2
28	hours.per.week	0.3	0.7	0.1
29	age	0.3	0.7	0.1
30	education.num	0.4	0.6	0.1
31	hours.per.week	0.4	0.6	0.1
32	age	0.4	0.6	0.1
33	education.num	0.5	0.5	0.1
34	hours.per.week	0.5	0.5	0.1
35	age	0.5	0.5	0.1
36	education.num	0.6	0.4	0.1
37	hours.per.week	0.6	0.4	0.1
38	age	0.6	0.4	0.1
39	education.num	0.7	0.3	0.1
40	hours.per.week	0.7	0.3	0.1
41	age	0.7	0.3	0.1
42	education.num	0.8	0.2	0.1
43	hours.per.week	0.8	0.2	0.1
44	age	0.8	0.2	0.1
45	education.num	0.9	0.1	0.1
46	hours.per.week	0.9	0.1	0.1
47	age	0.9	0.1	0.1

4.5.4 Estimated Analysis Results

Estimated methods, including Sturges' rule for bin width selection and curve fitting for selecting the best epsilon, were tested to validate the effectiveness of the hybrid metrics. These findings emphasized how crucial it is to balance privacy and usefulness when processing data automatically.

Finding the best possible balance between data protection and data utility is essential when it comes to data privacy. Discretizing continuous data attributes into bins is one of the process's key steps. An important factor in determining the outcome of privacy-utility trade-offs is the choice of bin width. Bin widths are typically manually set, which can add subjectivity and variability to the analysis. This section examines the automated bin width identification process that makes use of Sturges' rule, a technique that systematically establishes the number of bins according to the properties of the dataset.

A commonly used technique for determining the number of bins in histograms is Sturges' rule. It offers a simple and theoretically supported method by calculating the bin width based on the sample size's logarithm. This analysis seeks to standardize the evaluation of privacy-utility trade-offs across various data attributes and epsilon values by automating the binning process through the application of Sturges' rule.

A thorough understanding of the trade-offs between privacy and utility for various dataset attributes can be obtained from the results and visualizations produced by applying Sturges' rule to automatically calculate bin width. For this analysis, the sum of squared errors (SSE) and privacy gain were integrated with different weights to create a hybrid metric. A methodical approach to figuring out the ideal number of bins is provided by Sturges' rule, and the following analysis shows how this impacts the trade-off between privacy and utility.

In the estimated experiment, the number of bins was determined by applying Sturges' rule to the hybrid metric for the attribute `education.num` in the example shown in Figure 18. After that, a quadratic curve was fitted to this metric to determine the ideal epsilon value. The relationship between epsilon values and the hybrid metric for three distinct privacy and utility weight combinations is depicted in the figures.

The hybrid metric based on the curve fitting increases initially in the first chart, where the utility weight is 0.85 and the privacy weight is 0.15. It peaks at approximately 0.5 epsilon value, after which it declines. This pattern is closely followed by the quadratic fit, which suggests that the ideal epsilon is approximately 0.5. This peak indicates that, under these weight conditions, the balance between privacy and utility is maximized at this epsilon value.

A similar pattern can be seen in the second chart, which has a utility weight of 0.90 and a privacy weight of 0.10. The fitted curve shows that the hybrid metric increases, peaks, and then falls, with the peak occurring around an epsilon value of 0.6 and 0.7.

The third chart displays a decline in the hybrid metric after reaching a peak value, with the utility weight being 0.92 and the privacy weight being 0.08. Between 0.6 and 0.7 epsilon values are where the peak is seen. This is confirmed by the quadratic fit, which shows that the ideal epsilon value for this set of weights is between 0.6 and 0.7.

Together, Figure 18 show that various combinations of privacy and utility weights result in different optimal epsilon values, even for the same attribute. The application of Sturges' rule guarantees a methodical approach to bin selection, while curve fitting permits accurate identification of the ideal epsilon, which maximizes the sum of the metrics. This analysis emphasizes how crucial it is to modify epsilon in accordance with utility and privacy weightings to attain the optimal balance for various data attributes.

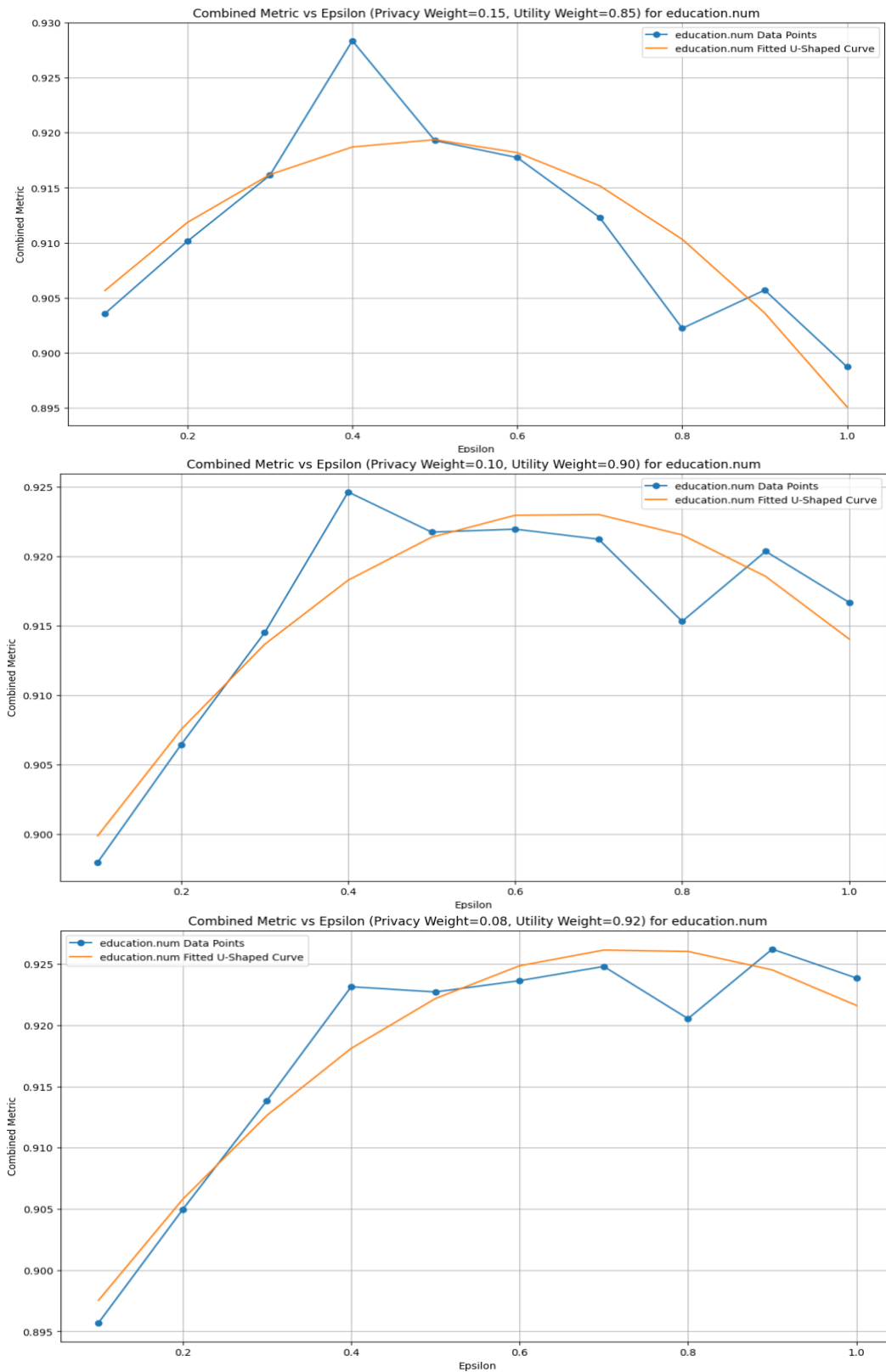


Figure 18: Combined Metric vs Epsilon for Different Privacy and Utility Weights in Estimated Experiment (Attribute: Education Number).

The best epsilon values for each attribute under various combinations of privacy and utility weights are summarized in Table 8. The goal of this analysis is to find the epsilon value that, while accounting for the trade-offs between privacy and utility, maximizes the hybrid metric. The table shows the optimal epsilon values under various combinations of privacy and utility weights for each attribute ('education.num', 'hours.per.week', and 'age'). Every combination result in a maximum value for the hybrid metric, which is the weighted sum of privacy gain and the inverse of SSE.

The optimal epsilon value for the 'education.num' attribute is consistently 1.0 at high utility weight values (e.g., 1.0, 0.99, 0.98, and 0.97). This pattern suggests that when utility is given priority, less noise is preferred. The ideal epsilon value falls as the privacy weight rises, indicating a greater noise tolerance to improve privacy. This trend indicates that as privacy concerns gain importance, it becomes necessary to balance them by adding more noise (lower epsilon). Similar trends are seen in the 'hours.per.week' attribute, where utility is highly prioritized and optimal epsilon values typically range between 0.759 and 0.781. The ideal epsilon value falls to 0.1 as the privacy weight rises, indicating a greater focus on maintaining privacy. This suggests that more noise is added to protect the data as the need for privacy increases. The ideal epsilon values for the 'age' attribute are mainly centered around 0.1, particularly in cases where the privacy weight is significant. On the other hand, the epsilon values can be higher, ranging from 0.657 to 0.780, when utility is given priority. This shows that the 'age' attribute can balance privacy and utility while tolerating higher noise levels. The different epsilon values show how flexible the 'age' attribute is to various trade-offs between privacy and utility.

Table 8: Best Epsilon Values for PG and SSE in Estimated Experiment.

	Column	Privacy Weight	Utility Weight	Best Epsilon	Hybrid Metric
0	education.num	0	1	1	0.997529
1	hours.per.week	0	1	0.781374	1.03488
2	age	0	1	0.657293	1.07847
3	education.num	0.01	0.99	1	0.994221
4	hours.per.week	0.01	0.99	0.780467	1.03381
5	age	0.01	0.99	0.1	0.979693
6	education.num	0.02	0.98	1	0.990321
7	hours.per.week	0.02	0.98	0.779436	1.03258
8	age	0.02	0.98	0.1	0.943612
9	education.num	0.03	0.97	1	0.985652
10	hours.per.week	0.03	0.97	0.778251	1.03118
11	age	0.03	0.97	0.1	0.933114
12	education.num	0.05	0.95	0.933676	0.97841
13	hours.per.week	0.05	0.95	0.775266	1.02767
14	age	0.05	0.95	0.1	0.925183
15	education.num	0.08	0.92	0.740003	0.936376
16	hours.per.week	0.08	0.92	0.768163	1.01943
17	age	0.08	0.92	0.1	0.920893
18	education.num	0.1	0.9	0.636285	0.854953
19	hours.per.week	0.1	0.9	0.759799	1.00992
20	age	0.1	0.9	0.1	0.919489
21	education.num	0.15	0.85	0.496657	0.838828
22	hours.per.week	0.15	0.85	0.642913	0.852243
23	age	0.15	0.85	0.1	0.917637
24	education.num	0.2	0.8	0.293391	0.868781
25	hours.per.week	0.2	0.8	0.1	0.991665
26	age	0.2	0.8	0.1	0.916719
27	education.num	0.3	0.7	0.101488	0.949807
28	hours.per.week	0.3	0.7	0.1	0.939824
29	age	0.3	0.7	0.1	0.915807
30	education.num	0.4	0.6	0.1	1.00237
31	hours.per.week	0.4	0.6	0.1	0.930453
32	age	0.4	0.6	0.1	0.915353
33	education.num	0.5	0.5	0.1	1.02902
34	hours.per.week	0.5	0.5	0.1	0.926529
35	age	0.5	0.5	0.1	0.915081
36	education.num	0.6	0.4	0.1	1.04513
37	hours.per.week	0.6	0.4	0.1	0.924371
38	age	0.6	0.4	0.1	0.9149
39	education.num	0.7	0.3	0.1	1.05593
40	hours.per.week	0.7	0.3	0.1	0.923007
41	age	0.7	0.3	0.1	0.914771
42	education.num	0.8	0.2	0.1	1.05662
43	hours.per.week	0.8	0.2	0.1	0.922066
44	age	0.8	0.2	0.1	0.914674
45	education.num	0.9	0.1	0.1	1.05508
46	hours.per.week	0.9	0.1	0.1	0.921379
47	age	0.9	0.1	0.1	0.914599

Table 9 presents the optimal epsilon values for every combination of weights, emphasizing the epsilon that optimizes the total metric covering all attributes. The ideal epsilon values when utility is highly valued are typically between 0.657 and 0.781. This

suggests a desire for less noisy settings to preserve high utility values and manageable levels of privacy. When the hybrid metric is maximized, it is evident that utility gains exceed the trade-off. The ideal epsilon value is approximately 0.1 for weight combinations where the privacy and utility weights are balanced, such as a 0.5 privacy weight and a 0.5 utility weight. This shows a more equitable trade-off where utility and privacy are given equal weight. The intermediate level of noise suggested by the epsilon value of 0.1 offers a sensible compromise between preserving data utility and safeguarding data privacy. When privacy is given top priority, the ideal epsilon value always decreases to 0.1. When the privacy weight is 0.7, 0.8, or 0.9, for example, weight combinations are observed where this indicates a higher tolerance for noise to improve privacy. Under these circumstances, the hybrid metric is still maximized, demonstrating that the elevated noise levels successfully preserve privacy while upholding levels of acceptable utility.

Ultimately, the examination of the optimal epsilon values for every attribute and overall weight combination shows how privacy and utility must be carefully balanced. The findings emphasize the necessity of carefully choosing epsilon values in accordance with the privacy and utility requirements, making sure that the hybrid metric is maximized to obtain the best possible trade-off.

Table 9: Overall Best Epsilon Values for PG and SS in Estimated Experience.

	Privacy Weight	Utility Weight	Best Epsilon	Hybrid Metric
0	0	1	0.657293	1.07847
1	0.01	0.99	0.780467	1.03381
2	0.02	0.98	0.779436	1.03258
3	0.03	0.97	0.778251	1.03118
4	0.05	0.95	0.775266	1.02767
5	0.08	0.92	0.768163	1.01943
6	0.1	0.9	0.759799	1.00992
7	0.15	0.85	0.1	0.917637
8	0.2	0.8	0.1	0.991665
9	0.3	0.7	0.101488	0.949807
10	0.4	0.6	0.1	1.00237
11	0.5	0.5	0.1	1.02902
12	0.6	0.4	0.1	1.04513
13	0.7	0.3	0.1	1.05593
14	0.8	0.2	0.1	1.05662
15	0.9	0.1	0.1	1.05508

4.5.5 Comparison of Empirical and Estimated Experiments

The analysis of empirical and estimated binning experiments demonstrates how important bin width selection is when weighing privacy versus utility trade-offs. Finding the ideal epsilon values to maximize a hybrid metric that combines privacy gain and accuracy was the aim of both strategies. While the estimated experiment used Sturges' rule to determine the bin widths systematically, the Empirical experiment specified bin widths based on the attributes.

The Empirical experiment revealed that the ideal epsilon values differed significantly depending on the privacy and utility weight combinations, indicating how sensitive each attribute was to the addition of noise. Table 7 shows that, for example, when the utility was prioritized (utility weight = 1), 'hours.per.week' exhibited a high sensitivity with an optimal epsilon of 0.6, while 'education.num' was more sensitive and needed an epsilon of 1. This suggests that various attributes have distinct noise-responses, and that custom bin specification can greatly affect these responses.

On the other hand, the estimated experiment that employed Sturges' rule offered a more uniform method for choosing bin widths. The ideal epsilon values were more constant across various privacy-utility weight combinations, as Table 8 illustrates. To maximize the hybrid metric, the 'hours.per.week' attribute consistently needed lower epsilon values, suggesting a preference for higher noise levels to preserve utility while protecting privacy. Comparably, "education.num" and "age" exhibited reliable trends, with ideal epsilon values frequently centered around 0.1 in cases where privacy weights were raised.

Comparing both experiments, when privacy is prioritized, the estimated approach using Sturges' rule typically produces lower epsilon values, indicating that a systematic binning method can result in a more conservative approach to privacy protection. For most attributes, the optimal epsilon values decrease to 0.1 in scenarios where privacy weights are greater than 0.5. This consistency demonstrates how well-automated binning works to standardize trade-offs between privacy and utility. Furthermore, the empirical experiment introduces subjectivity and variability because it relies on unique bin widths for each attribute, which may offer more nuanced control over the trade-offs. In contrast,

the estimated method ensures consistency across various datasets and analyses and provides a more objective and repeatable methodology.

In summary, while both estimated and custom binning techniques have advantages, the estimated strategy that applies Sturges' rule offers a more trustworthy framework for figuring out the ideal epsilon values. This makes it a useful tool for data anonymization and analysis by ensuring a balanced approach to privacy and utility across various attributes and weight combinations.

Chapter 5: Conclusion

In this thesis, we used differential privacy techniques to investigate the complex trade-off between data privacy and utility, with a particular emphasis on the use of Laplace noise. This work examined the effects of Laplace noise perturbation on the privacy and utility of different attributes in a dataset using two different hybrid metrics: one that combined Privacy Gain with SSE and another that combined Normalized Mutual Information (NMI) with Accuracy (1-SSE). The primary goal was to determine the ideal epsilon values (ϵ) under various weighting schemes that efficiently strike a balance between privacy and utility.

The goal for the hybrid metric combining Privacy Gain and Accuracy (1-SSE) was to maximize the hybrid metric. According to the analysis, higher epsilon values allowed for more significant data perturbation while still maintaining an acceptable level of utility, which generally resulted in a better balance between privacy gain and SSE. The ideal epsilon values, however, moved towards lower values as privacy weight increased, suggesting that more noise is required to improve privacy protection. There were some variations in sensitivity, but overall, this trend held for all attributes. 'Hours.per.week', for example, showed several peaks, indicating a more complex relationship with epsilon than the comparatively stable trends seen in 'education.num' and 'age'. The objective for the hybrid metric combining NMI and SSE was to minimize the hybrid metric. According to our research, smaller epsilon values—such as 0.1 and 0.2—offered better privacy protection but at the expense of higher SSE, which resulted in a notable reduction in data utility. When privacy is of the highest priority, these epsilon values are appropriate. Comparatively, situations where data utility is prioritized are better suited for higher epsilon values, such as 0.5 and 1.0, which provide better utility with lower SSE but less privacy protection.

A systematic method for discretizing continuous data attributes was made possible by automating the bin width selection process using Sturges' rule, which also improved consistency and reduced subjectivity. The estimated approach applied Sturges' rule consistently across all attributes, protecting a more uniform assessment of privacy-utility trade-offs than empirical experiments that depended on domain expertise.

Overall, this study emphasizes how crucial it is to carefully choose epsilon values depending on the privacy and utility requirements of the application. The results give useful insights into the trade-offs associated with data perturbation and provide helpful advice on how to optimize epsilon values to attain the appropriate trade-off between privacy and utility in a variety of data anonymization scenarios.

5.1 Managerial Implications

This research has important management implications. Data utility and privacy trade-offs need to be carefully considered by organizations that handle sensitive data. According to the results, choosing the right epsilon values for differential privacy mechanisms is essential to preserving the harmony between data accuracy and privacy protection.

Ensuring the safety of sensitive data while maintaining its analytical utility should be the top priority for managers and decision-makers. One effective technique for preserving privacy is to employ Laplace noise. Using this strategy can assist companies in adhering to privacy laws and fostering user confidence in data security.

Moreover, the knowledge gathered from this study can help organizations create data-sharing procedures and policies that protect privacy without sacrificing the usefulness of the data. Organizations can reduce privacy risks and improve their data governance procedures by taking a proactive approach to data privacy. Also, participants can find useful guidance on how to incorporate differential privacy algorithms into their data workflows within their organizations. The process includes explaining differential privacy principles to relevant parties, selecting suitable privacy parameters (such as epsilon values), and effectively incorporating the differential privacy technique into the existing data infrastructure and operations.

We highlight how important it is to continually enhance and adapt in the context of increasing risks to privacy, changing legal requirements, and speeding up technological advancements. Organizations need to be on alert for new developments in data security and protection. Reviewing and updating confidentiality measures and privacy policies

regularly is crucial. In addition, it is critical to cultivate a culture of privacy responsibility and awareness within the organization.

5.2 Research Implications

We make major contributions to the field of data privacy, both in terms of academic research and practical applications. By examining the effects of different epsilon values on privacy and utility metrics, it advances academic understanding of differential privacy approaches. The theoretical foundations of data privacy are strengthened by this analysis, which offers insightful information about how these methods can be successfully implemented in practical settings. From an analytical point of view, our study presents a new hybrid metric that integrates Sum Squared Error (SSE) with NMI, and accuracy with privacy gain providing a fresh way to assess privacy-preserving mechanisms. This methodological development not only makes the evaluation process better but also establishes a standard for future studies that create more thorough metrics for assessing how effective privacy techniques are.

The research findings have significant implications for businesses and data professionals. We offer useful guidance on choosing appropriate epsilon values for data perturbation by demonstrating the trade-offs between privacy and utility. With the help of this guidance, organizations can preserve the usability and quality of their data while still protecting it to the desired extent. Moreover, we present insightful recommendations for future study paths. It suggests looking into how data privacy has changed over time about data sharing policies and urges the creation of increasingly sophisticated technologies that protect privacy. These recommendations provide a road map for upcoming studies, assisting researchers in developing the field of data privacy. The study's conclusions also open up new avenues for investigation and advancement in privacy-preserving data analysis by providing a framework for the creation of fresh techniques and strategies that will improve data privacy and usefulness even more.

References

- Angrisani, A., Doosti, M., & Kashefi, E. (2023). A unifying framework for differentially private quantum algorithms. In *arXiv [quant-ph]*. <http://arxiv.org/abs/2307.04733>
- Arous, A., Guesmi, A., Hanif, M. A., Alouani, I., & Shafique, M. (2023). Exploring machine learning privacy/utility trade-off from a hyperparameters lens. *2023 International Joint Conference on Neural Networks (IJCNN)*.
- Avraam, D., Wilson, R., Butters, O., Burton, T., Nicolaides, C., Jones, E., Boyd, A., & Burton, P. (2021). Privacy preserving data visualizations. *EPJ Data Science*, *10*(1). <https://doi.org/10.1140/epjds/s13688-020-00257-4>
- Bhattacharjee, K., Chen, M., & Dasgupta, A. (2020). Privacy-preserving data visualization: Reflections on the state of the art and research opportunities. *Computer Graphics Forum: Journal of the European Association for Computer Graphics*, *39*(3), 675–692. <https://doi.org/10.1111/cgf.14032>
- Budiu, M., Thaker, P., Gopalan, P., Wieder, U., & Zaharia, M. (2022). Overlook: Differentially private exploratory visualization for big data. *JPC*, *12*(1). <https://doi.org/10.29012/jpc.779>
- Beyer, H. (1981). Tukey, John W.: Exploratory data analysis. Addison-Wesley publishing company reading, mass. — Menlo Park, cal., London, Amsterdam, Don Mills, Ontario, Sydney 1977, XVI, 688 S. Biometrical Journal, *23*(4), 413–414. <https://doi.org/10.1002/bimj.4710230408>
- Bugliesi, M. (2006). *Languages and Programming: 33rd International Colloquium*. Springer.
- Chou, J.-K., Bryan, C., & Ma, K.-L. (2017). Privacy preserving visualization for social network data with ontology information. *2017 IEEE Pacific Visualization Symposium (PacificVis)*.
- Chou, J.-K., Wang, Y., & Ma, K.-L. (2019). Privacy preserving visualization: A study on event sequence data. *Computer Graphics Forum: Journal of the European Association for Computer Graphics*, *38*(1), 340–355. <https://doi.org/10.1111/cgf.13535>
- Cormode, G., & Garofalakis, M. (2010). Histograms and wavelets on probabilistic data. *IEEE Transactions on Knowledge and Data Engineering*, *22*(8), 1142–1157. <https://doi.org/10.1109/tkde.2010.66>

- Cormode, G., Procopiuc, C. M., Srivastava, D., & Tran, T. (2018). Differentially Private Spatial Decompositions. *IEEE Transactions on Knowledge and Data Engineering*.
- Dasgupta, A., Chen, M., & Kosara, R. (2013). Measuring privacy and utility in privacy-preserving visualization: Measuring privacy in privacy-preserving visualization. *Comput Graph Forum*, 32(8), 35–47. <https://doi.org/10.1111/cgf.12142>
- Dhinakaran, D., Sankar, S. M. U., Selvaraj, D., & Raja, S. E. (2024). Privacy-preserving data in IoT-based cloud systems: A comprehensive survey with AI integration. In *arXiv [cs.CR]*. <http://arxiv.org/abs/2401.00794>
- Diaz, J., Meruvia-Pastor, O., & Vazquez, P.-P. (2018). Improving perception accuracy in bar charts with internal contrast and framing enhancements. *2018 22nd International Conference Information Visualisation (IV)*.
- Dwork, C. (2006). *Automata, languages and programming: 33rd international colloquium* (M. Bugliesi, Ed.). Springer.
- Dwork, C., & Roth, A. (2013). The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4), 211–407. <https://doi.org/10.1561/04000000042>
- Ehsan, H., Sharaf, M. A., & Chrysanthis, P. K. (2016). MuVE: Efficient multi-objective view recommendation for visual data exploration. *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, 731–742.
- Erlingsson, Ú., Pihur, V., & Korolova, A. (n.d.). Rounding-based approximation techniques for differential privacy. *Proceedings of the 35th International Conference on Automata, Languages and Programming (ICALP)*, 2014, 339–351.
- European Parliament, & Council. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation). *Official Journal of the European Union*, L119, 1-88.
- Feng, D., Huang, H., & Qu, H. (2020). *Differentially Private Visual Analytics: Challenges and Opportunities*. *IEEE Transactions on Visualization and Computer Graphics*, 26(1), 141-151. <https://doi.org/10.1109/TVCG.2019.2934675>

- Hay, M., Machanavajjhala, A., Miklau, G., Chen, Y., & Zhang, D. (2016). Principled evaluation of differentially private algorithms using DPBench. *Proceedings of the 2016 International Conference on Management of Data*.
- Ivanova, M., Trifonova, I., & Bogdanova, G. (2022). Privacy Preservation in eLearning: Exploration and Analysis. In *2022 20th International Conference on Information Technology Based Higher Education and Training (ITHET)* (pp. 1–8). IEEE.
- Jagadish, H., Koudas, N., Muthukrishnan, S., Poosala, V., Sevcik, K., & Suel, T. (1998). Optimal histograms with quality guarantees. *Very Large Data Bases Conference*. <https://www.semanticscholar.org/paper/0e12b1a390a8e6108cc8500acfadcb1f3409e535>
- Johnson, A., & Shmatikov, V. (2013). Privacy-preserving data exploration in genome-wide association studies. *International Conference on Knowledge Discovery & Data Mining. International Conference on Knowledge Discovery & Data Mining, 2013*, 1079–1087. <https://doi.org/10.1145/2487575.2487687>
- Kang, T., Kim, S., Sohn, J., & Awan, J. (2023). Differentially Private Topological Data Analysis. In *arXiv [stat.ML]*. <http://arxiv.org/abs/2305.03609>
- Le, T.-N., Lee, S.-M., Tran, P.-L., & Li, C.-S. (2023). Randomized response techniques: A systematic review from the pioneering work of Warner (1965) to the present. *Mathematics*, 11(7), 1718. <https://doi.org/10.3390/math11071718>
- Lee, H., Kim, S., Kim, J. W., & Chung, Y. D. (2017). Utility-preserving anonymization for health data publishing. *BMC Medical Informatics and Decision Making*, 17(1). <https://doi.org/10.1186/s12911-017-0499-0>
- Lindell, Y., & Pinkas, B. (2000). Privacy preserving data mining. In *Advances in Cryptology — CRYPTO 2000* (pp. 36–54). Springer Berlin Heidelberg.
- Ma, C., Yuan, L., Han, L., Ding, M., Bhaskar, R., & Li, J. (2023). Data level privacy preserving: A stochastic perturbation approach based on differential privacy. *IEEE Transactions on Knowledge and Data Engineering*, 35(4), 3619–3631. <https://doi.org/10.1109/tkde.2021.3137047>
- Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). L - diversity. *ACM Trans Knowl Discov Data*, 1(1). <https://doi.org/10.1145/1217299.1217302>

- Majeed, A., & Lee, S. (2021). Anonymization techniques for privacy preserving data publishing: A comprehensive survey. *IEEE Access: Practical Innovations, Open Solutions*, 9, 8512–8545. <https://doi.org/10.1109/access.2020.3045700>
- McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. Y. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (pp. 1273-1282).
- McSherry, F., & Talwar, K. (2007). Mechanism Design via Differential Privacy. *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, 94–103.
- Pramanik, M. I., Lau, R. Y. K., Hossain, M. S., Rahoman, M. M., Debnath, S. K., Rashed, M. G., & Uddin, M. Z. (2021). Privacy preserving big data analytics: A critical analysis of state-of-the-art. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, 11(1). <https://doi.org/10.1002/widm.1387>
- Roman, A.-S. (2023). Evaluating the privacy and utility of time-series data perturbation algorithms. *Mathematics*, 11(5), 1260. <https://doi.org/10.3390/math11051260>
- Ruggieri, S., Pedreschi, D., & Turini, F. (2010). Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data*, 4(2), 1–40. <https://doi.org/10.1145/1754428.1754432>
- Scott, D. W. (1979). "On optimal and data-based histograms." *Biometrika*, 66(3), 605-610. doi:10.1093/biomet/66.3.605
- Sparks, R., Carter, C., Donnelly, J. B., Okeefe, C. M., Duncan, J., Keighley, T., & Mcaullay, D. (2008). Remote access methods for exploratory data analysis and statistical modelling: Privacy-Preserving Analytics®. *Computer Methods and Programs in Biomedicine*, 91(3), 208–222.
- UAE Government. (2021). UAE Federal Data Protection Law No. 45 of 2021. Retrieved May 20, 2024, from <https://u.ae/en/about-the-uae/digital-uae/data/data-protection>
- UCI Machine Learning Repository. (2016). Adult census income [Data set]. University of California, Irvine. Retrieved from <https://archive.ics.uci.edu/ml/datasets/adult> on April 12, 2023.
- U.S. Department of Health and Human Services. (1996). Health Insurance Portability and Accountability Act of 1996 (HIPAA). Public Law 104-191.

- Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11, 2837–2854.
- Wagner, I., & Eckhoff, D. (2019). Technical privacy metrics: A systematic survey. *ACM Computing Surveys*, 51(3), 1–38. <https://doi.org/10.1145/3168389>
- Xiao, X., & Tao, Y. (2006). Personalized privacy preservation. *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*.
- Zhang, D., Sarvghad, A., & Miklau, G. (2021). Investigating visual analysis of differentially private data. *IEEE Transactions on Visualization and Computer Graphics*, 27(2), 1786–1796. <https://doi.org/10.1109/TVCG.2020.3030369>
- Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10), 931–934. <https://doi.org/10.1038/nmeth.3547>
- Zhou, J., Wang, X., Wong, J. K., Wang, H., Wang, Z., Yang, X., Yan, X., Feng, H., Qu, H., Ying, H., & Chen, W. (2023). DPVisCreator: Incorporating pattern constraints to privacy-preserving visualizations via Differential Privacy. *IEEE Transactions on Visualization and Computer Graphics*, 29(1), 809–819. <https://doi.org/10.1109/TVCG.2022.3209391>
- Zia, M. T., Khan, M. A., & El-Sayed, H. (2020). Application of differential privacy approach in healthcare data – A case study. *2020 14th International Conference on Innovations in Information Technology (IIT)*.



UAEU MASTER THESIS NO. 2024: 61

This thesis evaluates the balance between privacy and utility in data visualizations, using differential privacy techniques and hybrid metrics. It identifies optimal privacy settings to maximize data utility while ensuring privacy protection. This novel contribution helps to develop strategies that maximize data utility and privacy protection by offering a more nuanced understanding of how privacy-preserving techniques affect data visualization.

Sarah Hayi Alkaabi received her Master of Science in Information Technology Management from the Department of Information Systems and Security, College of Information Technology at the United Arab Emirates University, UAE. She received her Bachelor of Science in Information Security from the College of Information Technology, United Arab Emirates University, UAE.

www.uaeu.ac.ae

