Advancing Retail Data Science: Comprehensive Evaluation of Synthetic Data

Yu Xia [*], Chi-Hua Wang [†], Joshua Mabry [‡], Guang Cheng [§]

June 21, 2024

Abstract

The evaluation of synthetic data generation is crucial, especially in the retail sector where data accuracy is paramount. This paper introduces a comprehensive framework for assessing synthetic retail data, focusing on fidelity, utility, and privacy. Our approach differentiates between continuous and discrete data attributes, providing precise evaluation criteria. Fidelity is measured through stability and generalizability. Stability ensures synthetic data accurately replicates known data distributions, while generalizability confirms its robustness in novel scenarios. Utility is demonstrated through the synthetic data's effectiveness in critical retail tasks such as demand forecasting and dynamic pricing, proving its value in predictive analytics and strategic planning. Privacy is safeguarded using Differential Privacy, ensuring synthetic data maintains a perfect balance between resembling training and holdout datasets without compromising security. Our findings validate that this framework provides reliable and scalable evaluation for synthetic retail data. It ensures high fidelity, utility, and privacy, making it an essential tool for advancing retail data science. This framework meets the evolving needs of the retail industry with precision and confidence, paving the way for future advancements in synthetic data methodologies.

Key Words: Tabular Generative Models, Synthetic Data Fidelity, Machine Learning Utility, Synthetic Data Privacy, Retail Synthetic Data

[*]Machine Learning Engineer, Bain & Company, CA, 94158. Email: yu.xia@bain.com

[†]Postdoctoral Scholar, Department of Statistics & Data Science, UCLA, CA, 90095. Email: chihuawang@ucla.edu

[‡]Sr. Director of Data Science & ML Engineering, Bain & Company, CA, 94158. Email: joshua.mabry@bain.com

[§]Professor, Department of Statistics and Data Science, UCLA, CA, 90095. Email: guangcheng@ucla.edu

1 Introduction

In the rapidly evolving field of data science, the evaluation of synthetic data generation frameworks has become paramount, especially within the retail sector. This paper introduces a comprehensive framework for assessing synthetic retail data, focusing on three critical dimensions: fidelity, utility, and privacy. Our framework distinguishes between continuous and discrete attributes within retail datasets, providing clear methodologies for their evaluation.

Firstly, fidelity is evaluated through stability and generalizability [42]. Stability measures how well synthetic retail data replicates known data distributions, highlighting the robustness of models in familiar scenarios. Generalizability, on the other hand, assesses the performance of synthetic data in novel contexts, ensuring that the generated data can effectively extend beyond its training parameters. This is particularly important in retail, where market trends and consumer behavior can shift rapidly.

Secondly, the utility of synthetic retail data is scrutinized by its applicability to real-world tasks. In the retail sector, accurate demand forecasting and dynamic pricing [46, 27] are pivotal for operational efficiency and profitability. Our evaluation framework demonstrates how synthetic datasets can effectively support these core functions, making them indispensable for predictive analytics and strategic decision-making in retail.

Finally, privacy is assessed using Differential Privacy and related metrics. We compare the proximity of synthetic datasets to both training and holdout datasets to ensure balanced privacy guarantees. A well-balanced synthetic dataset should approximate both datasets equally, indicating robust privacy protection without compromising data utility. This aspect is critical in retail, where customer data privacy is a significant concern.

We apply our framework to evaluate generative AI models trained with the Complete Journey dataset [15]. Our results affirm that our evaluation framework provides a robust pipeline for large-scale assessments of synthetic retail data generation models. This framework not only ensures high fidelity and utility but also maintains stringent privacy standards. Consequently, it offers a solid foundation for future improvements in synthetic data generation and evaluation methodologies within the retail sector.

This paper concludes that with our framework, synthetic retail data can be reliably utilized for various applications, offering a scalable solution for the ever-growing demands of data privacy and utility in retail data science.

1.1 Background and Motivation

In the retail industry, the challenges of data privacy and availability are significant obstacles. Synthetic data, which is artificially generated rather than obtained from real-world events, provides a compelling solution to these issues. One primary challenge in retail is protecting customer privacy while leveraging data for analysis and decision-making. Synthetic data addresses this by mimicking real data without exposing sensitive customer information, and maintaining statistical properties and patterns found in actual data. This allows retailers to perform robust analyses and model training without risking data breaches or violating privacy regulations.

Moreover, obtaining large volumes of high-quality data can be difficult, especially when dealing with new products or services where historical data is sparse or non-existent. Public

datasets are notably smaller than standard industry datasets. They are generally collected under biased and often undisclosed marketing policies, and they lack many critical fields needed for accurate customer behavior modeling[14]. Moreover, business constraints and fairness concerns restrict the potential for aggressive experimentation across the marketing mix. Synthetic data generation overcomes these by creating abundant and varied datasets that reflect potential future scenarios or underrepresented cases. This capability is crucial for training machine learning models, which require large datasets to perform effectively. Additionally, synthetic data can help mitigate biases present in real data, leading to more fair and accurate models.

1.2 Objectives and Contributions

Developing a robust evaluation framework for synthetic data in retail is essential to ensure the validity and utility of the data. Without rigorous evaluation, synthetic data may fail to accurately reflect the complexities of real-world scenarios, leading to misleading insights and poor decision-making. A strong evaluation framework involves several critical components: assessing the statistical similarity between synthetic and real data, evaluating the impact on model performance, and ensuring that synthetic data preserves essential patterns and relationships. Thus, we propose a standardized evaluation framework for retail synthetic datasets from three aspects: fidelity, utility, and privacy.

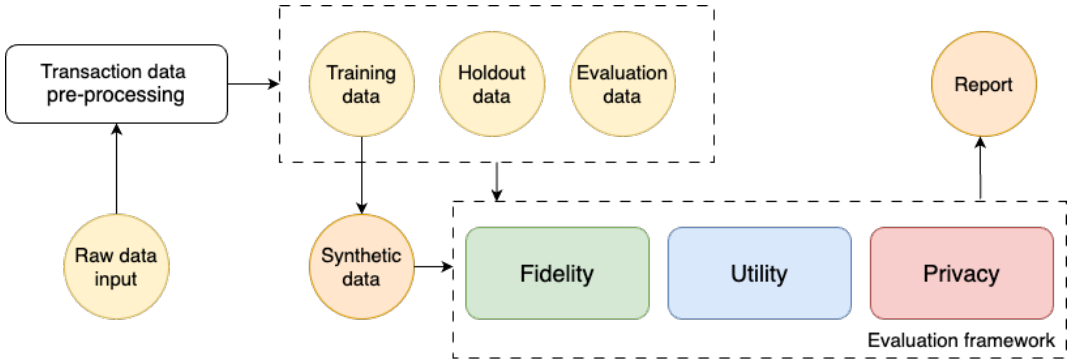

Figure 1: The framework diagram of our synthetic retail data evaluation pipeline. Section 3.1 explains the purpose and method to split transaction data. Section 3.2 defines detailed metrics for fidelity assessment, i.e. Wasserstein distance, Pearson correlation, etc. Section 3.3 defines the tasks for utility assessment, i.e. classification accuracy, product association, etc. Section3.4 explains the metrics for privacy assessment, i.e. distance to the closest record.

Such a framework (Figure 1) ensures that synthetic data is not only statistically similar to real data but also useful for practical applications in the retail sector. This process helps identify any discrepancies and areas where synthetic data may fall short, guiding improvements in data generation methods. Ultimately, a robust evaluation framework builds trust in synthetic data, making it a reliable resource for retailers. In this way, we ensure a safe and scalable way to generate high-quality synthetic data while maintaining privacy compliance.

2 Related Work

2.1 Existing Evaluation Frameworks

Several general frameworks have been proposed to gauge the efficacy of synthetic data previously. A sample-level metric framework evaluates generative models through fidelity and utility lenses, facilitating the identification of discrepancies and similarities between real and synthetic datasets [1]. Another methodology emphasizes auditing and generating synthetic data with controllable trust trade-offs, allowing customization based on specific requirements [5]. Further exploration of synthetic data generation discusses its benefits and limitations across various contexts, particularly in creating a practical approach for deployment [22].

Fidelity assessment ensures synthetic data retains the essential characteristics and patterns of real data. Previous work proposed a holdout-based empirical assessment method for mixed-type synthetic data, highlighting the importance of maintaining the statistical properties and variability inherent in the original dataset [32]. On the metric aspect, Sajjadi et al introduced a definition of precision and recall for distribution and quantified distribution similarity not just with one-dimensional score, like total variation [38].

Utility evaluation focuses on the synthetic data's performance in downstream tasks. Xu et al. established a basis of relevant utility theory in a statistical learning framework and introduced metrics of generalization and ranking of models trained on synthetic data [51]. It considers two utility metrics: generalization and ranking of models trained on synthetic data. There was also empirical work, for example, emphasizing generative model selection based on performance in fraud detection[13]. Additionally, Hsieh et al. (2024) adopted a data-centric perspective to improve both the fidelity and utility of synthetic credit card transaction time series [20]. Liu et al. explore utility in dynamic pricing models, demonstrating how synthetic data can support robust pricing strategies in fluctuating market conditions [28].

Privacy is a central concern in synthetic data generation. A formal framework for detecting data-copying in generative models ensures synthetic data does not replicate real data points [6], where the author also provides the requirement of minimum sample size for reliable detection. Meehan's work proposed a three-sample test to solve the same issue of data-copying in generative models [29]. BadGD addresses the vulnerabilities of gradient descent algorithms through strategic backdoor attacks to safeguard data privacy [45]. Furthermore, Chen et al. systematically summarize all approaches for differentially private data publishing to conduct reproducible downstream analysis while preserving data privacy [10]. Tools like TAPAS provide adversarial privacy auditing. A review of privacy measurement practices for tabular synthetic data includes a comprehensive list of privacy metrics [7], such as Differential Privacy, k-Anonymity, Plausible Deniability, etc.

However, to our knowledge, there is no empirical work conducting a full set of fidelity, utility, and privacy assessments on generative AI models with retail transaction data.

2.2 Synthetic Data in Retail

Synthetic data can transform the retail industry by enhancing various operational and analytical processes while ensuring customer privacy. It enables comprehensive customer analytics and segmentation without compromising personal data, aiding in the development of

targeted marketing strategies. In supply chain optimization, synthetic data simulates different scenarios to help forecast demand, optimize inventory, and improve logistics. Product recommendation systems can benefit from data augmentation with synthetic datasets for extensive training, ensuring accurate and relevant recommendations that enhance customer experience. Furthermore, synthetic data allows safe data sharing with third parties and partners under privacy regulations, facilitating collaborative projects and compliance checks without using real data, thus fostering innovation while safeguarding privacy.

Data practitioners have widely recognized the value of synthetic data for accelerating the development of AI systems and started to emphasize building generative models of the data in its raw tabular forms, instead of modeling features derived from transformed data [30, 22]. Researchers identify two different paths for synthetic data generation. One is the black-box style of privacy-preserving modeling techniques (such as Generative Adversarial Networks, Variational Autoencoders, and Bayesian Networks) [47]. For example, Athey generated synthetic data for the evaluation of causal effects estimators with Wasserstein Generative Adversarial Networks [4]. These privacy-preserving modeling techniques are powerful when sufficient historical data is available to learn an accurate data-generating process. However, when public retail datasets are rather scarce, the other style that needs domain knowledge in retail shines. Statisticians simplify the complexity of real-world data and specify structural causal models to generate synthetic data. For example, prior work simulated either category choice or the full life-cycle of customer shopping decisions based on a nested logit model [2, 48].

However, no matter which technique is used for synthetic data generation, there is currently no well-defined evaluation framework specifically designed to assess synthetic retail data, highlighting a crucial gap in ensuring data fidelity, utility, and privacy. The retail industry particularly requires effective data to conduct analysis, such as price optimization[11, 12], basket analysis [43, 36], customer lifetime value [35, 16], and demand forecasting [44, 17]. Given the complexity and importance of these tasks, careful evaluation of synthetic data is essential to preserve the quality of insights derived from these analyses, ensuring that strategic decisions are based on accurate and reliable information.

2.3 Complete Journey Dataset

The Dunnhumby Complete Journey Dataset [15] represents a comprehensive and meticulously curated collection of retail transaction data, providing deep insights into consumer purchase behaviors and patterns. Compiled from a vast array of shopping experiences, this dataset encompasses detailed records of customer interactions, including basket-level transaction details, promotional influences, and loyalty program participation. We utilize three main tables in our paper from the dataset: (1) transaction table, (2) customer demographics table, and (3) product hierarchy table. We merge these three tables with schema explained in Table 1 to build the raw data input to the proposed framework in Figure 1, and cover the pre-processing details in section 4.1.

Variable	Type	Description	Source
household_id	integer	Customer unique identifier	1 & 2
product_id	integer	Product unique identifier	1 & 3
day	integer	Transaction day	1
week	integer	Transaction week	1
quantity	integer	Purchased units	1
sales_value	float	Revenue	1
...	...	...	
age	string	Customer age group	2
household_size	string	Number of family members	2
...	...	...	
department	string	Product department	3
brand	string	"national" or "private"	3
...	...	...	

Table 1: Partial schema of merged Complete Journey dataset as an example, where (1) in the source column represent the transactions table, (2) for the customer demographics table, and (3) for the product hierarchy table. See Sec. 2.3 for full details.

3 The Evaluation Framework

In this session, we propose an empirical assessment framework to evaluate generative AI models for retail synthetic data generation. Our evaluation method is distinguished by its tripartite composition of synthetic data **Fidelity**, **Utility**, and **Privacy** metrics. The defining characteristic of this method is its adaptability and model-free nature, allowing it to be deployed independently of domain-specific knowledge or preconceived notions. Utilizing non-parametric measures, our data-centric evaluation provides a systematic review of an array of black-box synthetic data solutions, examining whether generated data is practical, safe, and broadly applicable. The objective underlining this methodology is to build transparency, enhance confidence in data generators, and further incentivize industries to leverage synthetic data for innovation in the modern data-driven world.

3.1 Train-Holdout-Eval Split

To robustly evaluate the generalizability of a data synthesizer, we employ a random split of the available records into three distinct datasets: a training dataset T, a holdout dataset H [32], and an additional evaluation dataset E, where the evaluation dataset E is only used for assessing model utility as a evaluation set (Figure 1). The training dataset T is exclusively used to train the synthesizer, while the holdout dataset H remains untouched during the synthetic data generation process. By exposing only the training dataset T to the synthesizer, we generate a synthetic dataset S of the same size as T. The isolated holdout dataset H can then serve as a benchmark to assess the synthesizer's ability to generalize beyond the data it was trained on.

To demonstrate the model generalizability, we compare the metrics obtained from both the holdout dataset H and the synthetic dataset S. If the holdout dataset H attains better metrics than the synthetic dataset S, the synthesizer has missed some underlying patterns presented in the data. Conversely, suppose the synthetic dataset S achieves superior metrics. In that case, this indicates potential overfitting by the synthesizer to the training dataset T. Ideally, we aim for the metrics from both datasets to be as close as possible, reflecting balanced generalization and reliable synthetic data generation.

3.2 Measuring Synthetic Data Fidelity

We treat holdout and synthetic datasets as separate data sources to evaluate fidelity metrics against the training dataset. The design of fidelity measurement is motivated by visualizing joint distributions and marginal distributions to discover patterns.

Similarity of Marginal Distribution One critical part of exploratory data analysis is to demonstrate the distribution of numerical features. This involves plotting histograms, density plots, and cumulative distribution functions to visualize how numerical data is spread across different ranges. A robust generative synthesizer must accurately learn and replicate these numerical distributions, ensuring that the synthetic data mirrors the real-world data in terms of central tendencies, variability, and distribution shape. Furthermore, we can also derive additional features from primitive columns and test their distribution similarities. This includes calculating ratios, differences, and other mathematical transformations to extract business insights. By assessing the distribution of these derived features, we can further evaluate the synthesizer's stability and robustness, ensuring that it captures intricate relationships and patterns within the data. For both primitive numerical features and derived numerical features, we report Wasserstein distance [37] to measure distribution similarities, where the small value indicates the synthetic dataset is closely attached to the real dataset.

Another important aspect is to check the distribution of categorical features. This involves visualizing the frequency of each category, cross-tabulations, and bar plots to understand the distribution of categorical data. A competent generative synthesizer must also learn these categorical distributions accurately. It should preserve the proportions and relationships among categories, ensuring that the synthetic data accurately represents the categorical structures observed in the real dataset. To quantify the degree of similarity, we compute the Jensen-Shannon distance [9] and expect a small value as an indicator of an excellent synthesizer.

Similarity of Joint Distribution Besides capturing the distribution of a single attribute, a synthesizer with high fidelity should also be able to identify multivariate combinations and relationships among the set of attributes, assessing how pairs of features interact and co-vary. We compute the Pearson correlation matrix for number-to-number interaction, Theil's U matrix for category-to-category interaction, and the correlation ratio matrix for number-to-category interactions. To verify if the synthesizer understands feature interactions and dependencies, we compute the L2 distance of flattened correlation arrays between the training dataset and the synthetic dataset or the holdout dataset. This step is vital for applications where the relationship between variables significantly impacts outcomes, such as customer segmentation and market basket analysis in the retail industry. Ensuring that these joint distributions are faithfully replicated in the synthetic data guarantees that the

model maintains the integrity of multivariate relationships, providing a more comprehensive and realistic representation of the underlying data structure.

3.3 Measuring Synthetic Data Utility

In evaluating the efficacy of generative models for retail synthetic data generation, a critical consideration is the preservation of Machine Learning (ML) utility. This step entails formulating a classification task $f : X \to Y$ using a predefined dataset, enabling a comprehensive assessment of how well-synthesized data can replicate real-world data's utility in predictive modeling. The evaluation framework is meticulously designed to ensure a robust comparison of model performance on both utility and generalizability.

To achieve this, we train machine learning models separately using the training dataset T, holdout dataset H, and synthetic dataset S. This approach allows us to systematically assess the performance of each model on the same evaluation set, referred to as evaluation dataset E. The model trained with T, f_T, provides a baseline for understanding performance metrics under standard conditions. Conversely, the model trained with H, f_H, offers insights into the model's behavior when exposed to unseen real-world data, thereby indicating its generalizability. Furthermore, the model trained with S, f_S, in this evaluative procedure enables a direct comparison of how well the generative models could replicate actual data characteristics and maintain predictive accuracy.

By testing f_T, f_H, f_S on evaluation dataset E and computing metrics like accuracy, F1, ROC, precision, and recall, we are able to systematically quantify and compare the utility of data generated by various generative AI models. This empirical assessment framework not only facilitates a granular understanding of each model's performance but also highlights the strengths and limitations of generative AI in capturing complex data patterns crucial for prediction tasks in the retail sector.

3.4 Measuring Synthetic Data Privacy

Privacy is a paramount concern in the realm of synthetic tabular data generation, primarily due to the sensitive nature of information often contained within retail datasets. The generation of synthetic data aims to mitigate the risk of disclosing private or proprietary information while still enabling valuable data-driven insights. To rigorously evaluate the privacy-preserving capabilities of generative AI models, we compute the Distance to Closest Record (DCR) , with L1 distance as the definition of distance between two records. Specifically, we assess the DCR from the synthetic data S to the training data T, and from the holdout data H to the training data T. The DCR quantifies the likelihood of synthetic data points being too similar to actual data points, thereby posing a privacy threat. A high DCR value indicates effective anonymization.

Additionally, we introduce a metric termed the **Closest Cluster Ratio (CCR)** further to scrutinize the privacy and generalizability of synthetic data. The CCR measures the proportion of synthetic data points that are closer to the training dataset compared to the holdout dataset, ranging from $[0, 1]$. Ideally, the values of CCR should be as low as possible, indicating that synthetic data points are not a close copy of the training dataset. A CCR

close to 1 signals an overfitting generative model, highlighting the necessity for continuous refinement in synthetic data generation techniques.

By combining DCR and CCR metrics, we can provide deep insights into how effectively synthetic data can protect sensitive information, thereby fostering trust and reliability in the deployment of generative AI solutions in real-world retail scenarios.

4 Evaluation Results

To demonstrate the proposed framework (Figure 1), we conducted an empirical assessment on the open-source Complete Journey dataset [15] of retail transactions from frequent customers in a retail grocery store (accessed from *completejourney-py* [1]). The dataset documents the purchasing patterns of more than two thousand households over a one-year period, who frequently shop at the retailer.

We examined 5 generative models to produce the synthetic datasets: GAN-based tabular generative models ((1) CTGAN[50], (2) AutoGAN) and Diffusion-based tabular generative models ((3) TabDDPM[23], (4) StasyAutoDiff[41], (5) TabAutoDiff[41]). Specifically, we implemented AutoGAN by preparing input features with AutoDiff [41] and training a GAN[18] using Torch [3]. Unless otherwise specified, models are cloned from the cited repositories and the training features are prepared according to encoding methods stated in the corresponding paper.

4.1 Data Description and Analysis

Data Preprocess The raw transaction data includes approximately 1.47 million transactions and a wide range of about 92,000 products. The dataset presents a detailed category hierarchy that includes product department, product category, and product type. It also offers comprehensive customer demographics, such as age, income, household size, and marital status. Key transaction information, like item quantity, transaction sales amount, and discounts, are documented, enabling the calculation of unit prices and discounts 1. We followed the same pre-process procedure shown in RetailSynth [48] to remove seasonality effects, by cleaning out unregistered customers, excluding transactions with non-positive transactions, de-duplicating the product catalog, removing infrequent products, and aggregating weekly transactions for each customer. This is a typical procedure for optimizing marketing spend, customer lifetime value calculation, etc. To further increase the effective data points for each customer, we clustered customers by their demographic information and ended up with a weekly retail transaction with about 251,000 records from 6000 products and 400 customer clusters.

Data Analysis To generate more customer- and product-level insights, we calculated derived features from the processed dataset, such as product purchase probability, store visit probability, basket size, etc. Figure 2 exhibits two numeric columns on the top row, showing the skewed distributions of native feature, quantity, and derived feature, basket size, in the

[1] https://pypi.org/project/completejourney-py/

real-world retail transaction dataset. The distribution of quantities purchased tends to be positively skewed because most customers typically buy products in small quantities. Bulk purchases are less frequent, leading to a long tail on the right side of the distribution. The "Basket Size" subplot shows the probability distribution of the total number of items in a customer's basket. The distribution is right-skewed, indicating that while most transactions have a lower total basket size, a few transactions involve significantly higher total purchases. This is typical in retail, where a small number of premium customers can drive a substantial portion of revenue.

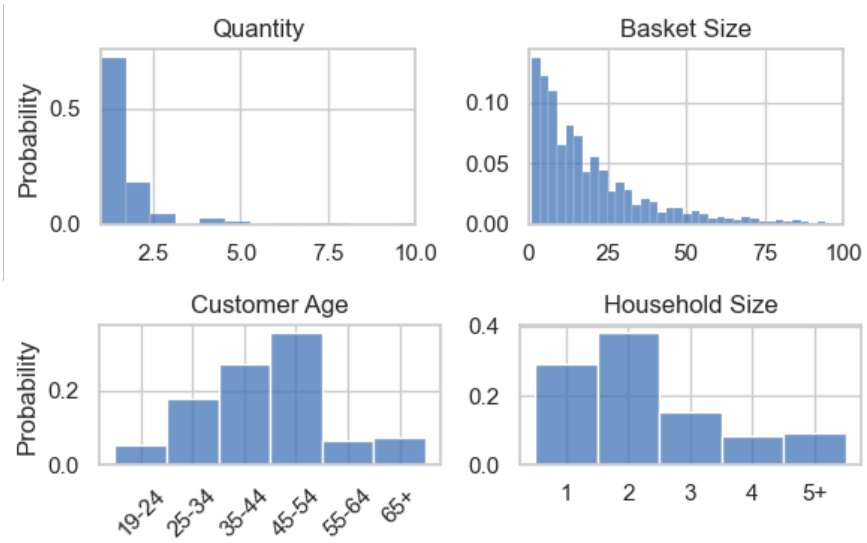

Figure 2: Selected univariate distributions for dataset "Complete Journey" illustrate diverse distributional patterns encountered in real-world datasets (see section 4.1).

Similarly, the bottom row details categorical distributions for "Customer Age", and "Household Size". Our dataset has a slight concentration of customers in the middle age groups (e.g., 35-44 and 45-54), suggesting that middle-aged consumers form a large portion of the customer base. However, younger (19-24, 25-34) age groups are also well-represented. Household size describes the number of members per household. The distribution peaks at household sizes of 2 and 3, indicating that most customers come from small to medium-sized households.

We also looked into multivariate combinations to explore relationships among the pair of columns. For example, Figure 3 demonstrates that as household size increases, the distribution of basket sizes progressively shifts to the right. In single-member households, purchases predominantly consist of smaller basket sizes. As household size grows from two to three and onward to five plus members, the distribution begins to show a higher density of larger basket sizes. This shift to the right indicates that larger households tend to buy more in a single transaction, reflecting their greater consumption needs.

Figure 4 shows a more comprehensive view on the Pearson correlation coefficient for numerical-numerical feature relationships, Theil's U statistic for assessing dependence between categorical-categorical features, and the correlation ratio for categorical-numerical feature associations. For example, a high positive correlation between manufacturer id (C3)

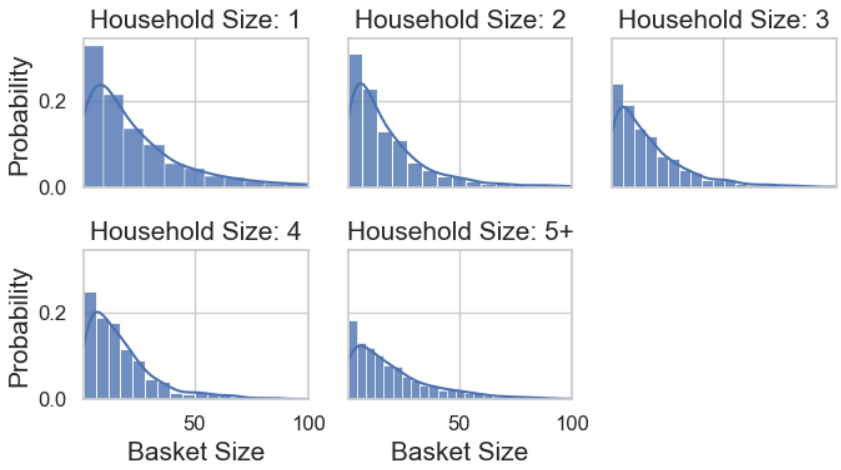

Figure 3: Marginal distribution of basket size by different household sizes. Customers with more family members in the household tend to buy more products in one visit (see section 4.1).

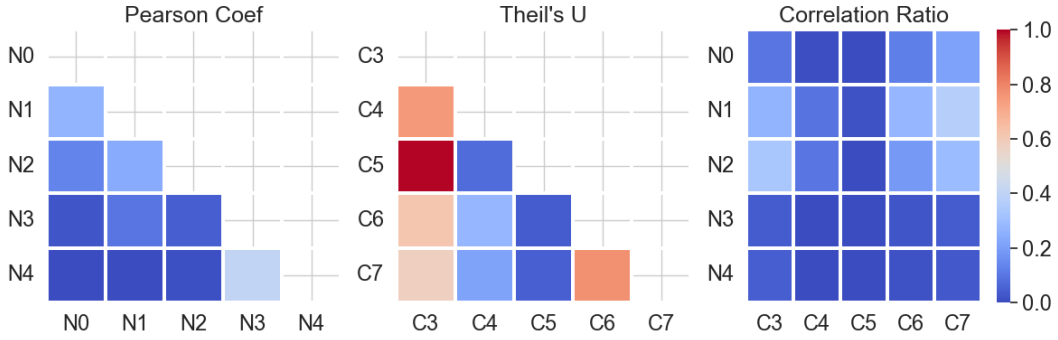

Figure 4: Correlation of selected numeric and categorical distributions for dataset "Complete Journey" illustrating contextual relationships observed in the real-world dataset (see Sec. 4.1).

and department (C4), see appendix A for a full list column mapping. Department, manufacturer, product category, product type, and package size are strongly associated with indicating a given product.

4.2 Synthetic Data Fidelity Assessment

Distributional similarity overview For the holdout dataset and synthetic datasets, we computed the average Wasserstein distance for numerical columns, the average Jensen-Shannon distance for categorical columns, and Euclidean distances of the Pearson correlation matrix, Theil's U matrix, and correlation ratio matrix against the training dataset. Table 2 presents a comprehensive evaluation of various generative AI models for retail synthetic data generation, highlighting the diverse strengths and weaknesses of each model.

Among the models evaluated, TabAutoDiff and CTGAN emerge as standout performers.

	Marginal			Joint	
	Num	Cat	Num-Num	Cat-Cat	Num-Cat
Holdout	0.04	0.38	0.45	0.04	0.04
CTGAN	2.24	0.46	**0.49**	4.06	**0.75**
AutoGAN	1646.88	0.41	2.13	8.15	4.28
TabDDPM	5.36	**0.38**	0.85	3.79	1.22
StasyAutoDiff	7.55	**0.38**	1.18	3.89	1.59
TabAutoDiff	**2.04**	0.42	0.63	**3.65**	0.81

Table 2: Fidelity metrics of similarities on marginal distributions and joint distributions, analyzed in 4.2. Different models have different strengths, with the best-performed model being highlighted for each metric. CTGAN and TabAutoDiff show more balanced performance from the fidelity aspect (see section 4.2).

TabAutoDiff demonstrates a consistently balanced performance across all metrics, excelling in capturing numerical marginal distributions and category-to-category joint distributions. This balanced prowess suggests TabAutoDiff's ability to effectively replicate retail datasets' complex, inherent patterns. CTGAN, on the other hand, excels particularly in learning number-to-number and number-to-category interactions. Both TabDDPM and SatAutoDiff models showed their strengths in learning categorical marginal distribution but failed to prove their intelligence in other distributions. We would not recommend AutoGAN from the fidelity perspective, because it did not stand out from any type of distribution examinations. . GAN models can lead to poor representation of the relationships and correlations among columns when the generator fails to capture the full diversity and complexity of the original dataset because of Mode Collapse. This can happen especially when the category columns present imbalanced distributions. However, CTGAN employs techniques like mode-specific normalization to stabilize the learning process.

Marginal Distribution Visualization We brought back features from Figure 2 and presented learned distributions from various generative models in Figure 5, plotting not only the feature distribution but also distributional differences between the real training data and the synthetic data. When it came to numerical features, quantity and basket size, TabAutoDiff and CTGAN were the only models that replicated the skewed distribution, though TabAutoDiff generated a small spike at the tail unexpectedly for the quantity distribution and CTGAN learned much fatter tails for both distributions. The diff plot proved These models' robustness again, where distributional difference histograms presented bars with height near 0. All models, except StasyAutoDiff and AutoGAN, were all capable to capture the right shape of category distributions, especially when the number of unique values in one category is low. However, if the dataset contains categorical columns with dramatic variation, we should expect a more concerning model performance.

Correlation Matrix Visualization To build a more informative presentation, we specifically provided a more detailed presentation on CTGAN, which performed the best in replicat-

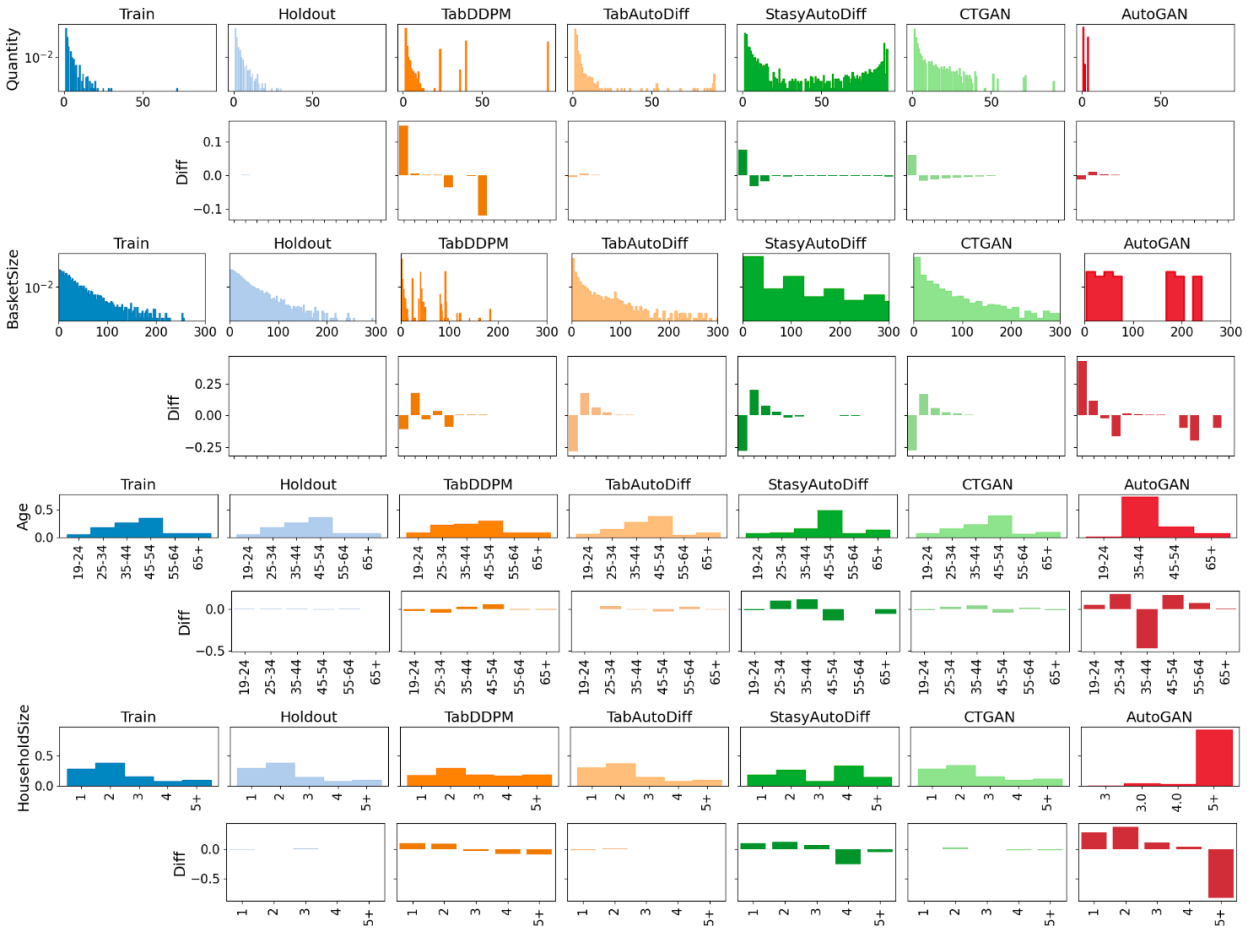

Figure 5: Distribution of feature columns from the training dataset, holdout dataset, and synthetic datasets, as well as the corresponding distribution difference to the one observed in the training dataset. The figure contains a primitive numerical column (Quantity), a derived numerical column (Basket Size), and primitive categorical columns (Age, Household Size), see 4.2. Synthetic data generated by TabAutoDiff demonstrates feature distributions that closely mirror the ones of the original training dataset.

ing joint distributions. Though the real dataset shows a weak correlation between features, CTGAN exhibits remarkable strength in capturing both numerical interactions and mixed-type interactions in Figure 6, highlighting its capability to generate coherent and realistic relationships within the data. This makes CTGAN a top choice for applications where high fidelity in interaction data is critical.

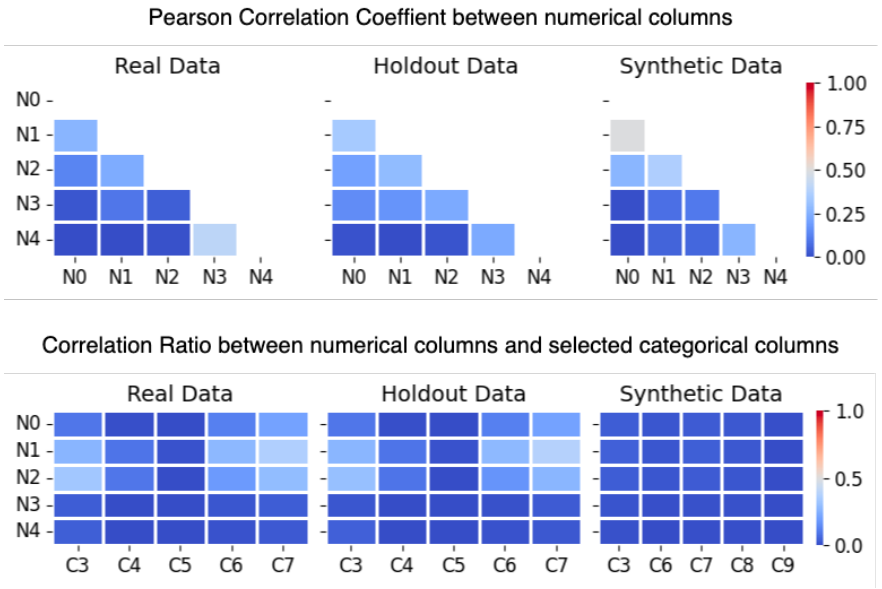

Figure 6: Heatmap of correlation metrics for Num-Num and Num-Cat interactions in the training, holdout, and CTGAN synthetic dataset. CTGAN model can replicate the feature interaction observed in the training dataset (see section 4.2)

Though we recommended TabAutoDiff and CTGAN from the fidelity perspective, all generative models have metrics larger than the corresponding values from the holdout dataset. This indicates that there are still hidden patterns in the training dataset that these models are not fully replicating. Synthetic data generation for retail is inherently challenging due to the high degree of heterogeneity and the dynamic nature of customer shopping preferences. Due to the complexity and challenges in synthetic data generation, generative models still have further headroom to capture the latent structure of retail datasets fully. The proposed evaluation framework is pivotal in this regard, as it provides a standardized approach to assess model stability and performance. By enabling a consistent comparison across different models and metrics, this framework aids in understanding the nuances of each model's strengths and areas for improvement.

4.3 Synthetic Data Utility Assessment

Classification Task To evaluate the model utility, we formulate two tasks. One is a classification task to identify premium customers who buy more products in one visit, predicting whether a customer will purchase more than 10 products based on their demographics and average unit price. We trained all classifiers supported by scikit-learn [31] and reported accuracy, F1, ROC, precision and recall of the model produces the highest accuracy, Bagging Classifier [8], in Table 3. Among the synthetic data models evaluated, TabAutoDiff emerges as the best-performing model for the classification task, indicating TabAutoDiff's superior performance in generating useful synthetic data that can effectively train classification models. When comparing the utility metrics of the synthetic data generated by TabAutoDiff to those of the train and holdout datasets, it is evident that the model trained with synthetic data achieved similar performance on all metrics, which indicates the capability of synthetic

data to generalize well to real data scenarios and serve as an effective proxy for real data. In this way, retailers can test marketing algorithms using synthetic data, eliminating the costs and risks of live A/B testing on real customers.

	Classification				
	Accuracy	F1	ROC	Precision	Recall
Train	0.65	0.62	0.67	0.52	0.76
Holdout	0.66	0.62	0.68	0.52	0.77
CTGAN	**0.68**	0.40	0.60	0.63	0.29
AutoGAN	0.38	0.53	0.50	0.37	**0.96**
TabDDPM	0.63	0.11	0.51	0.52	0.06
StasyAutoDiff	0.62	0.16	0.51	0.45	0.10
TabAutoDiff	**0.68**	**0.52**	**0.64**	**0.59**	0.47

Table 3: Utility metrics on a classification task. TabAutoDiff model achieves the best performance from the utility aspect as it scores high on accuracy, F1, ROC, and precision. See detailed description in Sec.4.3

Product Association Analysis The other task is to analyze product association in the training, holdout and synthetic datasets. We performed market basket analysis at the product level to see if there is a significant affinity for products to be purchased together using the Apriori algorithm [34], a popular method used in data mining for extracting such association rules. Table 4 presents the Lift metric, a measure of the likelihood that product B is bought when product A is bought, and the Conviction metric, which compares the probability that A appears without B if they were independent vs the actual frequency of A's appearance without B.

$$Confidence(A \to B) = P(A \cap B) \Big/ P(A) \tag{1}$$

$$Lift(A \to B) = Confidence(A \to B) \Big/ P(B) \tag{2}$$

$$Conviction(A \to B) = (1 - P(B))/(1 - Confidence(A \to B)) \tag{3}$$

where $P(A \cap B)$ is probability both products being purchased, and$P(A), P(B)$ is the individual probability of purchasing product A or B accordingly.

If Lift and Conviction much larger than 1, it means that product B is likely to be bought if product A is bought. In Table 4, both values for the synthetic datasets generated by AutoGAN and TabAutoDiff are significantly different from those of the Train and Holdout datasets. AutoGAN shows an exceptionally high lift and conviction values, indicating an overestimation of product pair occurrences, whereas TabAutoDiff's lift, although lower, still does not align closely with the real datasets. The synthetic data generated by CTGAN, TabDDPM, and StasyAutoDiff did not observe any frequently purchased product pairs, indicating a fundamental gap in learning the essential co-occurrence relationships of

products. This metric is invaluable for retailers as it helps identify product bundles, optimize placement strategies, and enhance cross-selling opportunities, thus leveraging consumer purchasing patterns to drive sales and customer satisfaction. By incorporating lift values, retailers can make data-driven decisions to refine their marketing and inventory strategies effectively. The evident difficulties reveal that existing generative models struggle in this aspect, which highlights the need for further refinement in replicating complex pairwise and higher-order relationships between products.

	Train	Holdout	AutoGAN	TabAutoDiff
Confidence	0.18	0.18	0.99	**0.26**
Lift	1.73	1.72	20.89	**4.83**
Conviction	1.10	1.09	inf	**1.30**

Table 4: Product association analysis for the listed synthetic data. Neither of the reported models can identify a similar product association rule observed in the training dataset. See analysis for each metric in Sec.4.3

4.4 Synthetic Data Privacy Assessment

Among the models we evaluated, TabAutoDiff showed a balanced performance in data privacy protection and model generalization. Table 5 presents privacy metrics for synthetic datasets generated by various models, focusing on the Distance to Closest Record (DCR) and the Closest Record Ratio (CCR). These metrics are crucial for assessing the privacy preservation capabilities of synthetic data, as they indicate how closely synthetic records resemble real training data points and the distribution balance, respectively.

Analyzing the DCR metric, the holdout set has the lowest DCR value, representing the benchmark distance within the real dataset. A larger DCR is preferred as it signifies that synthetic data points are not too close to any specific real training data points, thereby enhancing privacy. Among the models evaluated, TabAutoDiff demonstrates one of the higher DCR values. This indicates that its synthetic data maintains a significant privacy distance from the real data compared to other models, and stays at a lower risk of privacy leaking compared to the holdout dataset. TabAutoDiff also achieves the lowest CCR value, suggesting that the model did not overfit with the training data, thus providing good generalizability.

	DCR	CCR
Holdout	1.0	-
CTGAN	4.24	0.48
AutoGAN	8.82	0.51
TabDDPM	4.52	0.72
StasyAutoDiff	4.02	0.54
TabAutoDiff	**10.86**	**0.45**

Table 5: Privacy metrics of all tested models. TabAutoDiff stands out from the privacy aspect as it obtains the best performance in DCR and CCR. See detailed description in Sec.4.4

The retail industry is particularly cautious with customer data due to the sensitivity and privacy issues associated with handling such information. Strict regulations and the potential for reputational damage necessitate robust privacy preservation measures. The mixed performance of different models in terms of DCR and CCR highlights the need for a nuanced choice in synthetic data generation. While DCR is critical for ensuring individual data points are not too closely replicated, CCR helps ensure data generalizability, crucial for practical use in retail analytics. Thus, TabAutodiff is the best among evaluated models from the privacy perspective.

5 Discussions

Our comprehensive evaluation framework revealed distinct performances across various generative AI models in terms of fidelity, utility, and privacy for synthetic retail data. For fidelity, TabAutoDiff and CTGAN stood out, with TabAutoDiff demonstrating balanced performance across all metrics and CTGAN excelling in capturing joint distribution metrics, though all models showed room for improvement. In utility assessment, TabAutoDiff emerged as the top performer, effectively replicating the utility of real data in the classification task. However, none of the tested models demonstrate even a minimally acceptable performance in the product association analysis, indicating that further refinement of the model structure is necessary to achieve satisfactory results. For privacy, Distance to Closest Record (DCR) and Closest Cluster Ratio (CCR) metrics highlighted the models' ability to anonymize data effectively and balance generalization and identified TabAutoDiff again as the best one among the tested models. The proposed evaluation framework successfully highlighted the strengths and areas for improvement of each model, promoting the development of more effective and reliable synthetic data generation techniques for the retail sector.

Overall, our evaluation framework offers a standardized approach to assessing synthetic data generation models, facilitating consistent and transparent benchmarking. This can drive further research and development in the field, encouraging the creation of more sophisticated models capable of better capturing the complexities of retail data. To improve the evaluation framework, future research could focus on developing domain-specific metrics that capture the unique characteristics and complexities of various retail datasets, such as

transactional data, inventory data, etc. Expanding the diversity and size of the datasets used for evaluation, possibly through collaboration with industry partners or data-sharing initiatives, would enhance the robustness and generalizability of the evaluation framework's findings.

Anticipated advancements in synthetic data generation and evaluation include the development of more sophisticated generative models or LLM models capable of capturing higher-order dependencies and dynamic patterns presented in retail data. As these models evolve, they will not only improve in faithfully replicating the complexities of consumer behaviors but also in their ability to fill gaps where real-world data might be sparse or biased. Once the model reaches a mature level of learning all the patterns that retailers care about, we can confidently say that the generative model mirrors real customer purchase behavior. This validation would open up a multitude of applications, allowing the model to be used for inference such as demand forecasting and dynamic pricing. Furthermore, the robust nature of such advanced models could be leveraged within simulation environments to develop simulated A/B testing and test variations of a product or service in a controlled and cost-effective manner. A simulation environment with reliable generative models can also bridge the gap of reinforcement learning (RL) agents deployment aimed at personalized coupon-targeting strategies and optimizing customer engagement [49, 21]. By aligning synthetic data generation advancements with these application areas, businesses can not only gain deeper insights but also implement more dynamic and responsive retail strategies

References

[1] Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pages 290–306. PMLR, 2022.

[2] Rick L. Andrews, Imran S. Currim, and Peter S. H. Leeflang. A Comparison of Sales Response Predictions From Demand Models Applied to Store-Level versus Panel Data. *Journal of Business & Economic Statistics*, 29(2):319–326, April 2011.

[3] Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM, April 2024.

[4] Susan Athey, Raj Chetty, Guido Imbens, and Hyunseung Kang. Estimating Treatment Effects using Multiple Surrogates: The Role of the Surrogate Score and the Surrogate Index, February 2020. arXiv:1603.09326 [econ, stat].

[5] Brian Belgodere, Pierre Dognin, Adam Ivankay, Igor Melnyk, Youssef Mroueh, Aleksandra Mojsilovic, Jiri Navratil, Apoorva Nitsure, Inkit Padhi, Mattia Rigotti, et al. Auditing and generating synthetic data with controllable trust trade-offs. *arXiv preprint arXiv:2304.10819*, 2023.

[6] Robi Bhattacharjee, Sanjoy Dasgupta, and Kamalika Chaudhuri. Data-copying in generative models: a formal framework. In *International Conference on Machine Learning*, pages 2364–2396. PMLR, 2023.

[7] Alexander Theodorus Petrus Boudewijn, Andrea Filippo Ferraris, Daniele Panfilo, Vanessa Cocca, Sabrina Zinutti, Karel De Schepper, and Carlo Rossi Chauvenet. Privacy measurements in tabular synthetic data: State of the art and future research directions. In *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*, 2023.

[8] Leo Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

[9] Jop Briët and Peter Harremoës. Properties of classical and quantum jensen-shannon divergence. *Physical review A*, 79(5):052311, 2009.

[10] Dingfan Chen, Raouf Kerkouche, and Mario Fritz. A unified view of differentially private deep generative modeling. *arXiv preprint arXiv:2309.15696*, 2023.

[11] Xi Chen, Zachary Owen, Clark Pixton, and David Simchi-Levi. A statistical learning approach to personalization in revenue management. *Management Science*, 68(3):1923–1937, 2022.

[12] Xi Chen, David Simchi-Levi, and Yining Wang. Privacy-preserving dynamic personalized pricing with demand learning. *SSRN Electronic Journal*, 2020.

[13] Yinan Cheng, Chi-Hua Wang, Vamsi K Potluru, Tucker Balch, and Guang Cheng. Downstream task-oriented generative model selections on synthetic data training for fraud detection models. *arXiv preprint arXiv:2401.00974*, 2024.

[14] Alvise De Biasio, Andrea Montagna, Fabio Aiolli, and Nicolò Navarin. A systematic review of value-aware recommender systems. *Expert Systems with Applications*, 226:120131, September 2023.

[15] Dunnhumby. The Complete Journey, 2014.

[16] Yeliz Ekinci, Füsun Ulengin, and Nimet Uray. Using customer lifetime value to plan optimal promotions. *The Service Industries Journal*, 34(2):103–122, January 2014.

[17] Sebastian Gabel and Artem Timoshenko. Product choice with large assortments: A scalable deep-learning model. *Management Science*, 68(3):1808–1827, 2022.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[19] Florimond Houssiau, James Jordon, Samuel N Cohen, Owen Daniel, Andrew Elliott, James Geddes, Callum Mole, Camila Rangel-Smith, and Lukasz Szpruch. Tapas: a toolbox for adversarial privacy auditing of synthetic data. *arXiv preprint arXiv:2211.06550*, 2022.

[20] Din-Yin Hsieh, Chi-Hua Wang, and Guang Cheng. Improve fidelity and utility of synthetic credit card transaction time series from data-centric perspective. *arXiv preprint arXiv:2401.00965*, 2024.

[21] Eugene Ie, Chih-wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. RecSim: A Configurable Simulation Platform for Recommender Systems, September 2019. arXiv:1909.04847 [cs, stat].

[22] James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N. Cohen, and Adrian Weller. Synthetic Data – what, why and how?, May 2022. arXiv:2205.03257 [cs].

[23] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, pages 17564–17579. PMLR, 2023.

[24] Yuantong Li, Guang Cheng, and Xiaowu Dai. Two-sided competing matching recommendation markets with quota and complementary preferences constraints. In *Forty-first International Conference on Machine Learning*.

[25] Yuantong Li, Guang Cheng, and Xiaowu Dai. Dynamic online recommendation for two-sided market with bayesian incentive compatibility. *arXiv preprint arXiv:2406.04374*, 2024.

[26] Yuantong Li, Chi-hua Wang, Guang Cheng, and Will Wei Sun. Rate-optimal contextual online matching bandit. *arXiv preprint arXiv:2205.03699*, 2022.

[27] Po-Yi Liu, Chi-Hua Wang, and Henghsiu Tsai. Non-stationary dynamic pricing via actor-critic information-directed pricing. *arXiv preprint arXiv:2208.09372*, 2022.

[28] Yucong Liu, Chi-Hua Wang, and Guang Cheng. On the utility recovery incapability of neural net-based differential private tabular training data synthesizer under privacy deregulation. *arXiv preprint arXiv:2211.15809*, 2022.

[29] Casey Meehan, Kamalika Chaudhuri, and Sanjoy Dasgupta. A three sample hypothesis test for evaluating generative models. In *International Conference on Artificial Intelligence and Statistics*, pages 3546–3556. PMLR, 2020.

[30] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The Synthetic Data Vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410, October 2016.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[32] Michael Platzer and Thomas Reutterer. Holdout-based empirical assessment of mixed-type synthetic data. *Frontiers in big Data*, 4:679939, 2021.

[33] Pratik Ramprasad, Yuantong Li, Zhuoran Yang, Zhaoran Wang, Will Wei Sun, and Guang Cheng. Online bootstrap inference for policy evaluation in reinforcement learning. *Journal of the American Statistical Association*, 118(544):2901–2914, 2023.

[34] Sebastian Raschka. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *The Journal of Open Source Software*, 3(24), April 2018.

[35] Jaime Romero, Ralf van der Lans, and Berend Wierenga. A Partially Hidden Markov Model of Customer Dynamics for CLV Measurement. *Journal of Interactive Marketing*, 27(3):185–208, August 2013. Publisher: SAGE Publications.

[36] Francisco J. R. Ruiz, Susan Athey, and David M. Blei. Shopper: A probabilistic model of consumer choice with substitutes and complements. *The Annals of Applied Statistics*, 14(1), March 2020.

[37] Ludger Rüschendorf. The wasserstein distance and approximation theorems. *Probability Theory and Related Fields*, 70(1):117–129, 1985.

[38] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018.

[39] Marlesson R. O. Santana, Luckeciano C. Melo, Fernando H. F. Camargo, Bruno Brandão, Anderson Soares, Renan M. Oliveira, and Sandor Caetano. MARS-Gym: A Gym framework to model, train, and evaluate Recommender Systems for Marketplaces, September 2020. arXiv:2010.07035 [cs, stat].

[40] Sebastian Serth, Nikolai Podlesny, Marvin Bornstein, Jan Lindemann, Johanna Latt, Jan Selke, Rainer Schlosser, Martin Boissier, and Matthias Uflacker. An Interactive Platform to Simulate Dynamic Pricing Competition on Online Marketplaces. In *2017 IEEE 21st International Enterprise Distributed Object Computing Conference (EDOC)*, pages 61–66, October 2017. ISSN: 2325-6362.

[41] Namjoon Suh, Xiaofeng Lin, Din-Yin Hsieh, Merhdad Honarkhah, and Guang Cheng. Autodiff: combining auto-encoder and diffusion model for tabular data synthesizing, 2023.

[42] Lan Tao, Shirong Xu, Chi-Hua Wang, Namjoon Suh, and Guang Cheng. Discriminative estimation of total variation distance: A fidelity auditor for generative data. *arXiv preprint arXiv:2405.15337*, 2024.

[43] Luuk Van Maasakkers, Dennis Fok, and Bas Donkers. Next-basket prediction in a high-dimensional setting using gated recurrent units. *Expert Systems with Applications*, 212:118795, 2023.

[44] Mengting Wan, Di Wang, Matt Goldman, Matt Taddy, Justin Rao, Jie Liu, Dimitrios Lymberopoulos, and Julian McAuley. Modeling Consumer Preferences and Price Sensitivities from Large-Scale Grocery Shopping Transaction Logs. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1103–1112, Perth Australia, April 2017. International World Wide Web Conferences Steering Committee.

[45] Chi-Hua Wang and Guang Cheng. Badgd: A unified data-centric framework to identify gradient descent vulnerabilities. *arXiv preprint arXiv:2405.15979*, 2024.

[46] Chi-Hua Wang, Zhanyu Wang, Will Wei Sun, and Guang Cheng. Online regularization toward always-valid high-dimensional dynamic pricing. *Journal of the American Statistical Association*, pages 1–13, 2023.

[47] Harrison Wilde, Jack Jewson, Sebastian Vollmer, and Chris Holmes. Foundations of Bayesian Learning from Synthetic Data. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 541–549. PMLR, March 2021. ISSN: 2640-3498.

[48] Yu Xia, Ali Arian, Sriram Narayanamoorthy, and Joshua Mabry. RetailSynth: Synthetic Data Generation for Retail AI Systems Evaluation, December 2023. arXiv:2312.14095 [cs, econ, stat].

[49] Yu Xia, Sriram Narayanamoorthy, Zhengyuan Zhou, and Joshua Mabry. Simulation-based benchmarking of reinforcement learning agents for personalized retail promotions, 2024.

[50] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. In *Advances in Neural Information Processing Systems*, 2019.

[51] Shirong Xu, Will Wei Sun, and Guang Cheng. Utility theory of synthetic data generation. *arXiv preprint arXiv:2305.10015*, 2023.

A Column name mapping

The complete journey data has relative long names for each column. For a succinct presentation of the correlation heatmap in figure 4, we created a column mapping to label numerical columns and categorical columns.

Raw name	Type	Label
product_id	categorical	C1
household_id	categorical	C2
week	categorical	C3
manufacturer_id	categorical	C4
department	categorical	C5
brand	categorical	C6
product_category	categorical	C7
product_type	categorical	C8
package_size	categorical	C9
age	categorical	C10
homeownership	categorical	C11
marital_status	categorical	C12
household_size	categorical	C13
household_comp	categorical	C14
kids_count	categorical	C15
quantity	numerical	N1
sales_value	numerical	N2
retail_disc	numerical	N3
coupon_disc	numerical	N4
coupon_match_disc	numerical	N5
unit_price	numerical	N6

Table 6: Column name mapping to the shortened label.