THE
BEHAVIOURAL
INSIGHTS
TEAM

EEF Synthetic Data Project

High fidelity synthetic data generation feasibility study technical report

Authors: Clare Brennan, Cath Sleeman, Bobby Stuijfzand

Table of Contents

Appendices: Detailed Results

List of tables and figures

Executive summary

This feasibility study examines whether the Education Endowment Foundation (EEF) is able to produce high fidelity synthetic data for their archive of 128 Randomised Controlled Trial (RCT) datasets. Synthetic data is a privacy-preserving technique, which is "a new copy of a data set that is generated at random but made to follow the structure and some of the patterns of the original data set."[1] High fidelity synthetic datasets mimic both the original distribution of the variables, and some relationships between them, but do not contain the original observations.

To assess the feasibility of producing high fidelity synthetic data for EEF's RCT archive, we tested three different methods for producing synthetic data on three representative datasets. The three methods were a Conditional Classification and Regression Tree (Conditional CART), a Maximum Spanning Tree (MST) and a Generative Adversarial Network (GAN). These methods, and their pros and cons, are described in detail within the report.

We tested each method on a dataset representative of one of EEF's larger datasets (67,200 observations), a mid-sized dataset (9,900 observations) and a small dataset (1,200 observations). We found that for both of the larger datasets, both the Conditional CART and MST methods produced reasonable results, showing that for many of EEF's RCTs it is technically feasible to produce high fidelity synthetic data. The GAN produced poor results for the mid size dataset, and could not be trained in a reasonable time on the larger dataset. There was no method that produced reasonable results for the small dataset (1,200 observations), indicating that high-fidelity synthetic data may not be feasible for smaller RCTs.

Whilst the technical results are promising, there are risks to producing high fidelity synthetic data. We recommend that EEF considers processes to minimise these risks when deciding on whether or not to produce and publish high-fidelity synthetic data; in particular the risk that individual data is revealed either through incorrect production of synthetic data, or a reconstruction attack. To minimise these risks, there are a number of steps EEF can take, most importantly controlling access to high fidelity synthetic datasets (for example, releasing only to named, accredited researchers outside of a secure environment). This could strike an appropriate balance between reducing risk, and improving access. These risks and potential mitigations are explored in detail in the final section of this report.

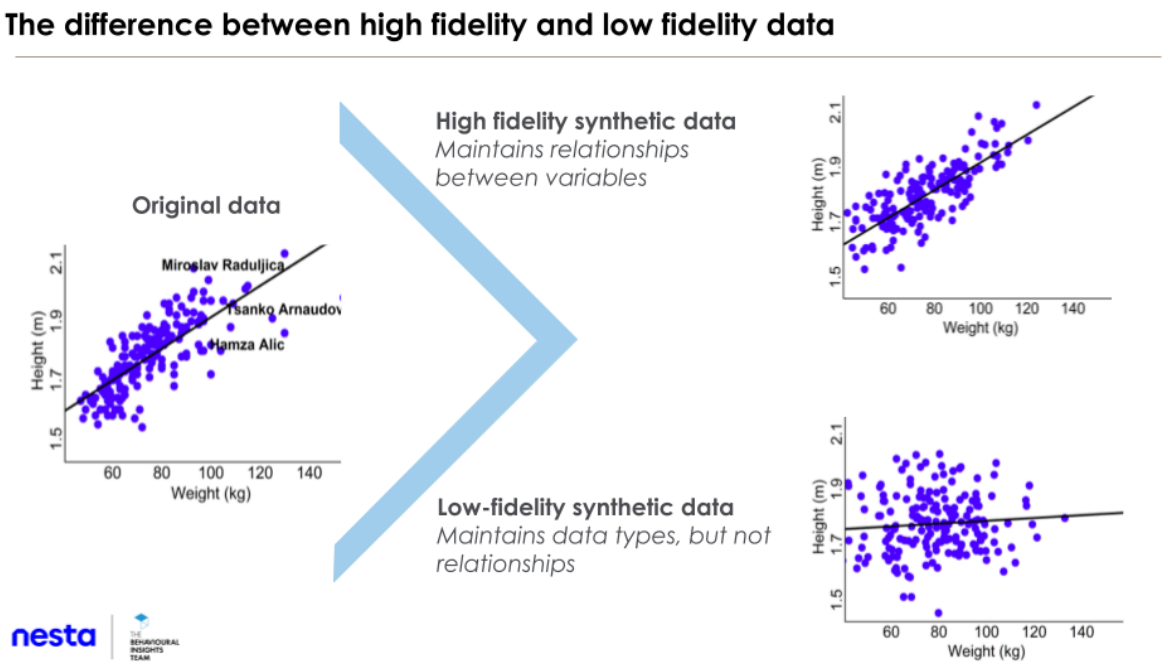
[1]Calcraft, P., Thomas, I., Maglicic, M., & Sutherland, A. (2021, December 14). *Accelerating public policy research with synthetic data*. Report investigates how synthetic data can be used in government, p2. https://www.adruk.org/news-publications/news-blogs/report-investigates-how-synthetic-data-can-be-used-in-government/

Introduction

This feasibility study examines whether the Education Endowment Foundation (EEF) is able to produce high fidelity synthetic data for their database of 128 (to date) Randomised Controlled Trial (RCT) datasets. It complements additional work, undertaken in parallel, to produce low fidelity synthetic versions of all datasets in the EEF archive. The report builds on previous work from the Behavioral Insights Team (BIT), undertaken for Administrative Data Research UK (ADR UK), called Accelerating public policy research with synthetic data[2].

"[S]ynthetic data is a new copy of a data set that is generated at random but made to follow the structure and some of the patterns of the original data set."[3] As seen in Figure 1 below, high fidelity synthetic data is data which aims to mimic both the variable distributions, and the relationships between variables.

Figure 1: The difference between high fidelity and low fidelity synthetic data



To assess the overall feasibility of producing high fidelity synthetic data for EEF's datasets, we've considered three elements:

- **Fidelity:** Is it possible to produce datasets that closely mimic the distribution of variables, and relationships between the variables, in representative EEF datasets?

[2] Ibid
[3] Ibid, p2

- **Privacy:** Can these datasets be produced whilst ensuring the privacy of participants in EEF-funded RCTs (by reducing the risk of 'leakage' of real observations into the synthetic datasets below an acceptable threshold)?
- **Utility:** What is the additional utility of making high fidelity synthetic data available to researchers, above and beyond the planned low fidelity datasets?

As is broadly noted in the literature available on synthetic data, there is a trade-off between the fidelity and privacy of a synthetic dataset. Datasets with higher fidelity (more of the original relationships preserved) will also have a greater risk that true data is 'leaked' into the synthetic dataset.[4] A core aim of this feasibility study was therefore to examine how this privacy-fidelity trade-off could be managed using EEF data, and if a sufficiently high *fidelity* could be achieved whilst maintaining *privacy* guarantees, such that the synthetic data will be useful to researchers (maximising the third element, *utility*).

[4] Jordan, J., Houssiau, F., Cherubin, G., Cohen, S., Szpruch, L., Bottarelli, M., Maple, C., & Weller, A. (2022, May 22). Synthetic Data - what, why and how? arxgiv>Computer Science> Machine Learning. https://arxiv.org/abs/2205.03257;

Wilde, H., Lawrence, M., Houssiau, F., & Daniel, O. (2019, February 21). Synthetic data for public good. Data Science for Public Good. Retrieved April 24, 2024, from https://datasciencecampus.ons.gov.uk/projects/synthetic-data-for-public-good/

Methodology

To generate the final datasets, we undertook a six-step process of:

Data selection: We selected three datasets, representative of EEF data in terms of their sample size and variable types. Whilst the initial plan for this study was to produce high-fidelity synthetic datasets on EEF's original data within the SRS, for computational, access, and software reasons (see Box 1) we ultimately conducted the study on datasets that were similar to EEF datasets, but were available outside SRS.

Data cleaning: We cleaned the data to remove redundant columns, ensure correct data types, and manage missing values. See Appendix A for more detail.

Method selection: We conducted a rapid review of available methods to determine the methods to test. Based on the desktop scan, we selected three methods to test on the data - Classification and Regression Trees (Conditional CART), Maximum Spanning Tree (MST), and Generative Adversarial Networks (GAN). A description of each method, and the python libraries used to generate data, is below. We generated high-fidelity synthetic data using each of these methods.

Model benchmarking: For each model, we benchmarked the dataset's fidelity, utility, and privacy using metrics from ONS' *synthgauge* package. We selected the best performing method for each dataset individually to produce the final synthetic dataset.

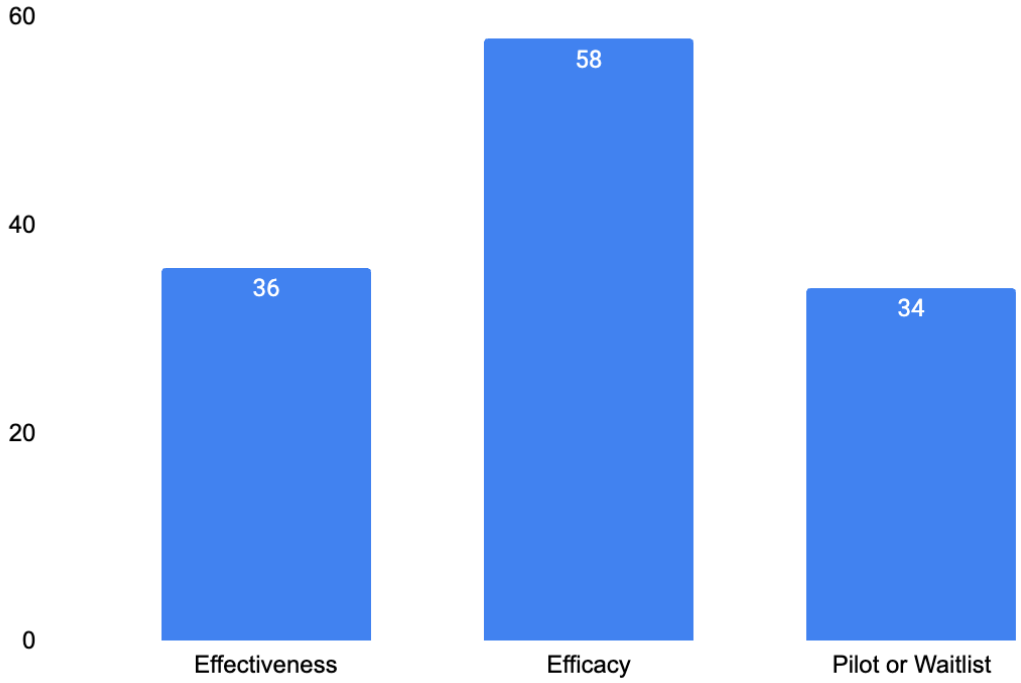
Model optimisation: Using the method that performed best on the initial benchmarks, we generated high-fidelity synthetic data. We optimised the model, including training hyper-parameters and choosing the optimal level of differential privacy.

Final evaluation: We re-benchmarked the final optimised model, using the same metrics, to give a final measure of its fidelity, utility, and privacy (as the model's performance may have changed after it was optimised). We used these metrics to place the datasets within the ONS' draft framework for privacy.

Data selection

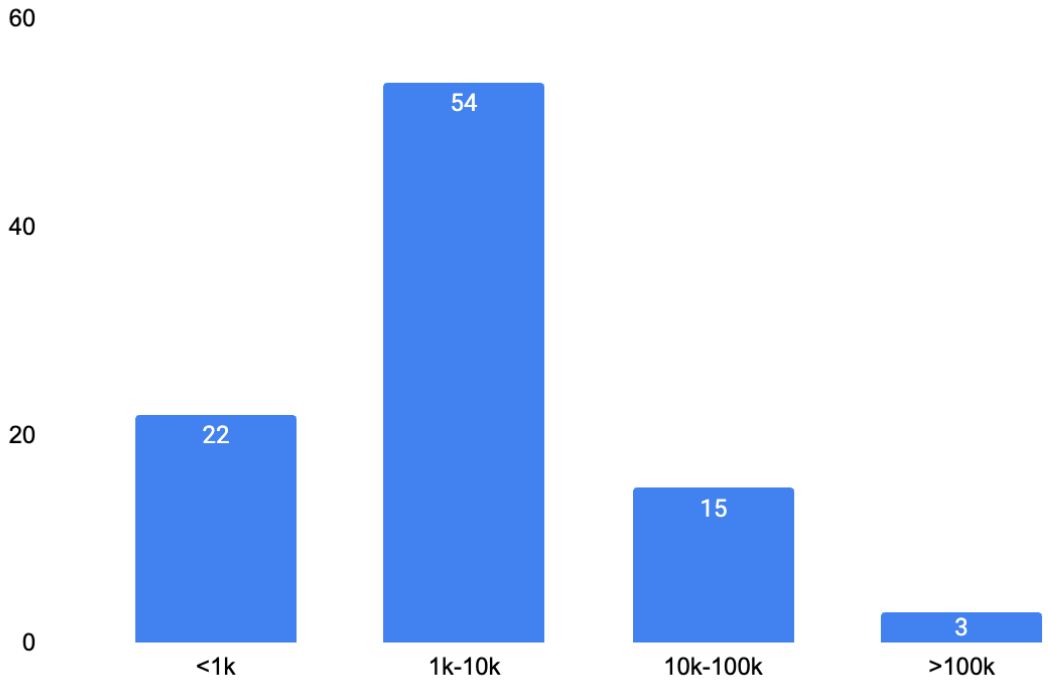
The initial goal for this study was to produce high fidelity data on 3-5 representative datasets within the EEF archive, stored in the ONS Secure Research Service (SRS). We were able to access the SRS and carry out an initial analysis of the archive (presented below), but due to ongoing issues accessing and working within the SRS, we ultimately conducted the study on alternative data, working outside the SRS (see Box 1 for a detailed explanation for the change in approach).

Figure 2: Number of EEF data sets by project type[5]



[5] This chart was created in the Office for National Statistics Secure Research Service using data from ONS and other owners and does not imply the endorsement of the ONS or other data owners.

Figure 3: Count of EEF data sets by sample size[6]



To create datasets that would be representative of EEF data, we considered aspects of the data known to affect the efficacy of various methods for synthetic data generation:[7]

- Study size (# of observations)
- Data types (continuous, categorical, boolean)
- Proportion of missing data
- Presence of imbalanced classes within the dataset

[6] "This chart was created in the Office for National Statistics Secure Research Service using data from ONS and other owners and does not imply the endorsement of the ONS or other data owners."

[7] Pathare, A., Mangrulkar, R., Suvarna, K., Parekh, A., Thakur, G., & Gawade, A. (2023). Comparison of tabular synthetic data generation techniques using propensity and cluster log metric. International Journal of Information Management Data Insights, 100177(3), 1-12. Elsevier. 10.1016/j.jjimei.2023.100177

Table 1: Summary of datasets used for feasibility study

Name	Sample size	# of variables by type
Maths Champions	1,200	Dichotomous: 14 variables Continuous: 6 variables
Magic breakfast	9,900	Dichotomous: 8 variables Continuous numeric: 12 variables Categorical (more than 2 categories): 2 variables
Inclusive learning	67, 200	Dichotomous: 18 variables Continuous: 18 variables

Box 1: Why work outside the SRS?

This project took place from November 2023 - June 2024, just as the ONS Secure Research Service (SRS) was beginning to transfer to the Integrated Data Service (IDS), a cloud-based secure environment. It was challenging to do data science work (in python) in the SRS, due to the manual management of the python environment and dependencies, and limited computational power:

- All changes to python packages (including installing additional dependencies, upgrading packages, or installing new packages) had to be carried out by the SRS's IT team, which typically meant a wait of ~5 days for each individual change or dependency issue. The SRS runs older versions of standard python packages, limiting their functionality and potentially limiting the additional packages that can be installed.
- Models took ~10x longer to run within the SRS environment than with similar data on a local CPU outside the environment.

The combination of these issues meant that, had we continued work within the SRS, the determining factor in our results and analysis would have been what was possible within the computational and python environment constraints of the SRS.

We decided to work outside the SRS, on datasets that resembled EEF datasets in the ways that matter for high-fidelity synthetic data generation. We expect this will provide a more realistic view of EEF's *future ability* to produce high fidelity synthetic data (working

within the IDS cloud environment), as the IDS is not likely to have the same issues of package management or computational power.[8]

Data cleaning

To effectively generate synthetic data, we cleaned the datasets to remove data that doesn't need to be synthetically generated (like columns giving text descriptions of other columns, columns which are completely empty, or columns which have the same vale in every cell). Then we standardised the format of each column and replaced missing values with values that maintained the overall statistical profile of the dataset. More detail on approach taken can be found in Appendix A.

Method selection

There are a variety of methods and approaches to generating high fidelity synthetic data, each with different strengths and weaknesses. To narrow down the methods to test on EEF data, we scanned the available literature on existing synthetic data methods.Table 2 shows the methods ultimately selected. For each method, we considered:

1. Has this been shown to produce high utility, high fidelity synthetic data for datasets that are similar to our datasets?
2. Is it possible to protect privacy using this method? Specifically, does it enable differential privacy? (see Box 2 for an explanation of this term)
3. Can it be produced with relatively small computational power (on a CPU)?
4. Is there an open-source python package that will enable us to implement this method, from a reputable source?
5. Can it be easily explained?
6. Across the methods selected to test, do we have a good range of approaches represented?

There are two caveats to our approach for selecting methods to test. Firstly, for the purposes of the feasibility study, we were not resourced to conduct a comprehensive literature review of all possible synthetic data generation methods. Therefore, we cannot definitively claim that the methods selected to test were the best possible methods. The methods ultimately tested represent a selection of common methods, shown to have worked on similar datasets, and fulfilling the practical considerations listed above.

Secondly, all methods considered assume access to the original datasets when producing high fidelity synthetic data. It is theoretically possible to produce high fidelity synthetic data from summary statistics alone, using graph-based approaches like the Bayesian or MST

[8] UK Office for National Statistics. (2024, 04 24). *About the Integrated Data Service*. Available Tooling. https://integrateddataservice.gov.uk/available-tooling

methods described below.[9] However, doing so was beyond the scope of this initial feasibility study.

[9] Wilde, H., Lawrence, M., Houssiau, F., & Daniel, O. (2023). *Synthesising the linked 2011 Census and deaths dataset while preserving its confidentiality*. Data Science for the Public Good.

Table 2: Methods selected for testing *If a method was found to be unsuitable on one criteria, it was dropped from investigation on further criteria.*

	Conditional classification and regression trees (Conditional CART)	**Proportional Fitting Classification and Regression Trees[10]**	**MST**	**Bayesian**	**Gaussian Copula**	**Generative Adversarial Networks**	**Variational Auto-Encoders**	**Recurrent neural networks**
Selected to test on the dataset?	**Yes**	No	**Yes**	No	No	**Yes**	No	No
Has this been shown to produce high utility, high fidelity synthetic data for data similar to our datasets? [11]	**Good performance on all data types**	**Good performance on all data types**	**Good performance, but continuous variables are transformed into discrete values**	Moderate performance on categorical data, poorer on continuous data	Good performance on categorical data, poorer on continuous data	**Moderate performance on all data types**	Poor performance on similar datasets (more suited to image generation)	Good performance on continuous data
Does it enable DP (see Box 2 for an explanation of this term)?	**No**	**Yes**	**Yes**	Yes	Yes	**Yes**	Not investigated	Yes
Can it be produced with relatively small computational power (on a CPU)?	**Yes**	**Yes**	**Yes**	Yes	Yes	**Yes, with small datasets**	Not investigated	No
Is there an open-source python package to easily implement this method, from a reputable source?	***Hazy implementation of synthpop***	No - future work can use R synthpop package	***pgm-private***	*synthcity*	No	***synthcity***	Not investigated	Not investigated
Can it be easily explained?	**Yes**	**Yes**	**Yes**	Yes	Not investigated	**No**	Not investigated	Not investigated

[10] Raab, G. (2022). Utility and Disclosure Risk for Differentially Private Synthetic Categorical Data. *Privacy in Statistical Databases*. https://arxiv.org/pdf/2206.01362

[11] Pathare, A., Mangrulkar, R., Suvarna, K., Parekh, A., Thakur, G., & Gawade, A. (2023). Comparison of tabular synthetic data generation techniques using propensity and cluster log metric. *International Journal of Information Management Data Insights*, *100177*(3), 1-12. Elsevier. 10.1016/j.jjimei.2023.100177; Bowen, C., & Snoke, J. (2020). Comparative Study of Differentially Private Synthetic Data Algorithms from the NIST PSCR Differential Privacy Synthetic Data Challenge. *arXiv*. arXiv:1911.12704v3.

Box 2: What is Differential Privacy (DP)?[12]

Differential privacy, or DP, is a formal mathematical technique for generating synthetic data in a way that strengthens its privacy guarantee (minimising the risk that an individual's true data can be reconstructed from the synthetic data).

As the BIT report *Accelerating public policy research with synthetic data*[13] explains:

"While there are some variations depending on the precise method used, typically when differential privacy is used to create synthetic data, some additional randomness (calibrated noise) is injected into the data. This is done in such a way that aggregate statistical properties are largely unaffected while individual records are perturbed, so that the probability of reproducing an individual's data or inferring their membership of the data set is reduced.

In [a typical] differential privacy scheme, the degree of privacy granted is controlled by a parameter ε (epsilon) [the privacy budget]. There is a trade-off between privacy and the utility of the data set: as ε is made smaller it will be less likely that an individual record can be inferred to be a member of the data set [stronger privacy], but the noise introduced will increase and may render the synthetic data useless for analysis [lower fidelity]."

Method 1: Classification and regression trees (Conditional CART)

The Conditional CART method is a tree-based approach to generating synthetic data. It is one of the most widely used methods for generating synthetic data across the UK government, including for the Scottish Longitudinal study (see Box 2).[14] For the Conditional CART method, the true data is split, with each node as a variable (eg, age, gender). It's split again and again, creating a tree, and the tree ends with data points that are very similar to one another in each leaf (see Figure 2 for a visual example). Then, synthetic data is generated by picking random values (from a reasonable range) for a variable, and using it to predict the data points in the leaves - so we have data that resembles the original data statistically, but that has come from random points. The exact points to split the variables on

[12] For more detail on DP, see this helpful explainer video from the US Census Bureau: https://www.youtube.com/watch?v=pT19VwBAqKA

[13] Calcraft, P., Thomas, I., Maglicic, M., & Sutherland, A. (2021, December 14). *Accelerating public policy research with synthetic data*. Behavioural Insights Team | ADR. https://www.adruk.org/fileadmin/uploads/adruk/Documents/Accelerating_public_policy_research_with_synthetic_data_December_2021.pdf

[14] Nowok, B., Raab, G., & Dibben, C. (2017). Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R. *Statistical Journal of the IAOS*, *33*, 785-796. ResearchGate. 10.3233/SJI-150153

are chosen to minimise the total squared distance between the true data points and predicted data points,[15] mimicking the conditional distribution of the variables.

The Conditional CART method has been found in the literature to perform consistently well, particularly for small datasets.[16] In comparison to graph-based methods, it is particularly good at capturing non-linear relationships.[17] It is very quick to run, relatively easy to explain, and has a strong history of effective and safe use to provide high fidelity synthetic data. Because this study used only python (due to time and resource constraints), we were only able to implement the Conditional CART method using a commercially-developed package in beta version, *Hazy Synthpop*[18], which offers only older CART methods, based on conditional fitting, without a differential privacy guarantee. Future work could use the popular R package *synthpop*[19], which offers a newer CART method, based on iterative proportional fitting, with the option of a differential privacy guarantee.[20]

For the purposes of this study, we used the python implementation of the *synthpop* package from Hazy Ltd, available as an open source package on GitHub.[21]

Box 3: Synthetic data for the Scottish Longitudinal Study

The Scottish Longitudinal Study (SLS) has provided synthetic data to researchers since 2015. Their process of generating and distributing synthetic data demonstrates how layers of both technical and process controls can be used to effectively balance fidelity and privacy.

The data is produced with the Conditional CART method, using the popular R package *synthpop*. Whilst there is no DP mechanism for this method, the generated data is checked to ensure that there is no privacy loss (no replication of true data within the synthetic dataset), and that the risk of reconstruction is below acceptable levels.

---

[15] Pathare et al, 2023

[16] Pathare et al, 2023

[17] Nowok, B., Raab, G., & Dibben, C. (2017). Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R. *Statistical Journal of the IAOS*, *33*, 785-796. ResearchGate. 10.3233/SJI-150153

[18] Hazy Ltd. (2017, November 9). synthpop README. Github: hazy/synthpop. Retrieved April 25, 2024, from https://github.com/hazy/synthpop?tab=readme-ov-file#readme

[19] (n.d.). synthpop - Generate synthetic data. Retrieved April 25, 2024, from https://synthpop.org.uk/

[20] Raab, G. (2022). Utility and Disclosure Risk for Differentially Private Synthetic Categorical Data. *Privacy in Statistical Databases*. https://arxiv.org/pdf/2206.01362

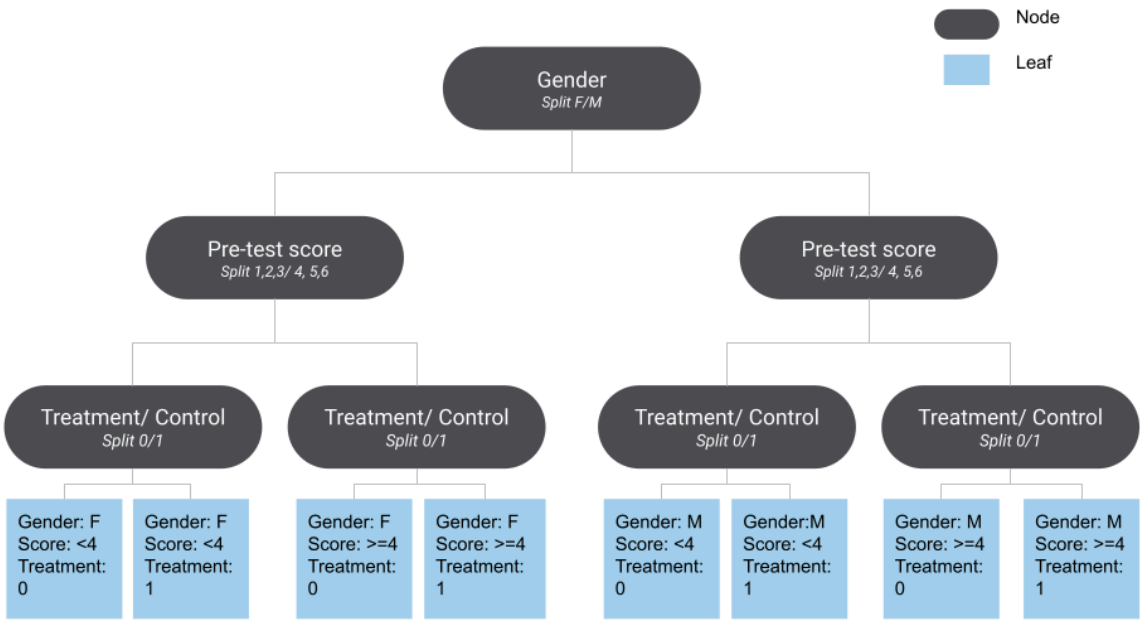
[21] Hazy Ltd. (2017, November 9). *synthpop README*. Github: hazy/synthpop. Retrieved April 25, 2024, from https://github.com/hazy/synthpop?tab=readme-ov-file#readme

A ‘spine’ of synthetic data, with 60,000 observations and five variables (age, marital status, general health, religion, approximated social grade), measured at two points in time, is made freely available on the SYLLS website.[22]

Researchers also have the option to request a bespoke synthetic dataset from the SLS. Bespoke data can be provided outside the safe setting, but only to named researchers, and only after they have undertaken safe researcher training. The data can be used to refine code, but not to publish results (reflecting a balance struck between fidelity and privacy).[23]

This approach has been extended over time, and synthetic data spines are now available for the England and Wales Longitudinal study, and the Northern Ireland Longitudinal study.[24]

Figure 4: Illustrative example of a Classification and Regression Tree[25]



[22] *Synthetic LS data*. (2017, November 9). CALLS -HUB. Retrieved April 26, 2024

[23] How to access SLS synthetic data. (2017, November 9). Scottish Longitudinal Study Development & Support Unit. Retrieved April 26, 2024, from https://sls.lscs.ac.uk/guides-resources/synthetic-data/how-to-access-sls-synthetic-data/

[24] *Synthetic LS data*. (2017, November 9). CALLS -HUB. Retrieved April 26, 2024

[25] Freie Universitat Berlin. (n.d.). *Random Forest*. Blogs@FU-Berlin. Retrieved April 25, 2024, from https://blogs.fu-berlin.de/reseda/random-forest

Method 2 : Maximum Spanning Tree (MST)

Like a range of similar graph-based methods available, MST finds the marginal distribution for variables in the sample (the probability that a value occurs for a single variable, independent of the other variables). Then, it adds some noise to the marginal distribution (this is the step that introduces DP; and the amount of noise is controlled by the privacy threshold ϵ). Finally, it generates synthetic data that preserves the marginal distributions.[26]

MST has been shown to work well on similar data to EEF's (it placed first in the final round of the NIST competition of 2018, a global competition for synthetic data generation methods).[27] More recently, it was the method selected by the ONS's Data Science Campus to trial synthetic data generation of the 2011 Census.[28] In addition to being a high-fidelity method which enables DP, it is auditable (as opposed to a black box model). It is also possible to allocate the privacy budget to specific variables using the MST method, enabling greater preservation of some relationships, and more noise in others. This is a particularly helpful feature for RCT data like EEF's, where we want to preserve the treatment effect, but may be more willing to introduce noise into explanatory variables. The key disadvantage of the MST method is that it requires discrete values (like all graph-based methods), so any continuous variables would need to be converted to discrete categories before generating synthetic data. [29] However, the method is able to take such a high number of discrete values that data can be 'binned' (grouped together) into very small ranges. As a result, MST was still an effective method for generating synthetic data from the UK census, with a number of continuous variables including ages and dates, with 55 million observations each.[30]

MST was ultimately selected for use over a Bayesian method (mentioned in the original bid for this project) as it has a similar underlying methodology, the same DP guarantee, but has previously created higher fidelity data with lower computational costs (placing first in the 2018 NIST challenge, where a comparable bayesian model placed third).[31]

Box 4: Generating synthetic data for the 2011 UK Census[32]

In 2023 the Office for National Statistics (ONS), in collaboration with the Alan Turing Institute, published a report on creating a synthetic dataset for the 2011 Census and deaths register. The original dataset contains all responses to the 2011 England and Wales census, along with any deaths registered up to the census date in 2019 (a total of 55 million rows and 60 columns).

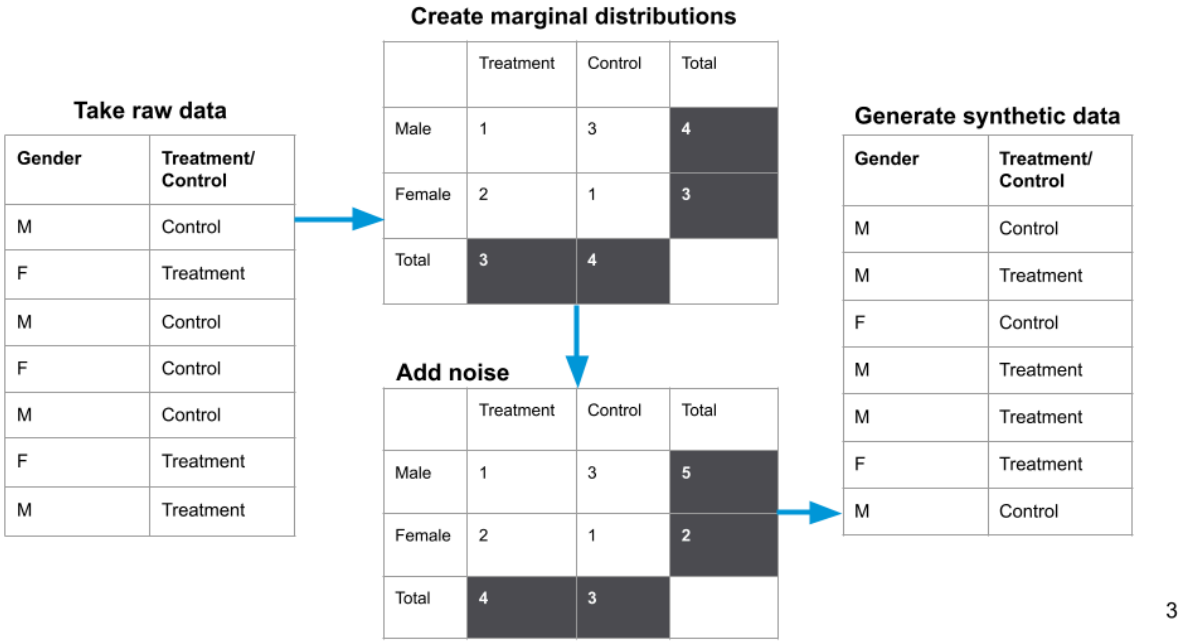
The ONS team selected the MST method for their project, using an adaptation of the python *private-pgm* package. Along with its high fidelity (shown by strong performance in

[26] McKenna, R., Miklau, G., & Sheldon, D. (2021). *Winning the NIST contest: a scalable and general approach to differentially private synthetic data*. http://arxiv.org/abs/2108.04978
[27] Bowen & Snoke, 2020
[28] Wilde et al., 2023
[29] Bowen & Snoke, 2020
[30] Wilde et al., 2023
[31] Bowen & Snoke, 2020
[32] Wilde et al., 2023

the 2018 NIST competition) and capacity for DP, the report highlights the explainability of the method as a key strength. Importantly, because the model uses marginal distribution tables as the basis for data generation (a relatively basic statistical concept), the team were able to engage stakeholders within the UK civil service in key design decisions when implementing the method. For example, the report highlights how the team were able to consult stakeholders on the parameters for the privacy budget and agree values informed by the ONS' existing statistical disclosure control guidelines, something their stakeholders are familiar with as a longstanding and widely understood privacy control.

The technical report's discussion of both the strengths and limitations of the method were very helpful when designing this feasibility study, notwithstanding the considerably larger dataset used in the ONS' study. The full report is available on the ONS website, and work is ongoing at the time of writing to publish the synthetic data within the Integrated Data Service.

Figure 5: Illustrative example of the graph-based synthetic data generation process



3

Method 3: Generative Adversarial Networks (GANs)

Generative Adversarial Models are a type of neural network that are increasingly popular for generating synthetic data.[33] They work by creating two models - a 'generator' and 'discriminator'. The 'generator' creates a synthetic dataset from the original by learning features of the data. The 'discriminator' then samples from the combined dataset and attempts to accurately distinguish between the 'true' and 'synthetic' data. The errors from that

[33]Pathare et al., 2023

classification task are used to improve the performance of the generator. This cycle is repeated until the discriminator cannot distinguish between data points from the synthetic or original datasets.

The advantages of GANs are that they can produce very high fidelity data with strong DP guarantees. However, they are data-hungry, and typically require datasets with millions of observations (unlike EEF data, where datasets are in the tens or hundreds of thousands of observations). In some recent comparisons, alternative approaches like CART and MST have in fact produced higher fidelity data for smaller datasets (like EEF's).[34] Moreover, they are computationally expensive to train, and as black box models, are difficult to audit or explain.

[34] Bowen & Snoke, 2020

Model benchmarking

After generating synthetic data using the methods described above, we used a selection of metrics, available through the ONS *synthgauge* package,[35] to benchmark each method. Whilst there is no single agreed approach for how to evaluate synthetic data, the metrics available through the *synthgauge* package are widely used, comprehensive, and easy to both implement and explain. We selected a range of metrics to capture fidelity, utility and privacy from within the package, outlined in Table 3. We note that, were further work to be completed using the *R* package synthpop, appropriate metrics would be available within the package to evaluate utility, fidelity and privacy.[36]

Table 3: Benchmarking model performance - metrics used

	Metric	Explanation
Fidelity	Correlation mean squared difference: Mean-squared difference in pairwise Pearson correlation coefficients (continuous data)	Measures how well relationships between continuous variables have been preserved, by taking the difference in correlations between variables between the synthetic and original data. *(Minimise towards 0)*
	pMSE-ratio: Ratio of the propensity score mean square error	Measures the accuracy of probabilistic forecasts by evaluating both the squared difference between predicted and actual values and the uncertainty in the predictions.
	MAD: Mean absolute difference of feature densities	Measures how well the distribution of individual variables has been preserved, by taking the difference in the measure of their distribution between synthetic and original data. *(Minimise towards 0)*
Utility	Classification error: Classification comparison (difference in precision, recall, F1 scores on a classification task)	Uses the synthetic data to train a classifier, and then tests the difference in accuracy when predicting on real data. *(Minimise towards 0)*
Privacy	Nearest neighbours: Minimum distance nearest neighbour	A check to ensure that outliers have not been replicated in the synthetic data *(Maximise)*
	TCAP score: Target Correct Attributional Probability Score	The risk that a target variable can be generated given a key variable *(Minimise towards 0)*
	Sample overlap: Proportion of real data found in synthetic data	A straightforward check that no real observations are contained in the synthetic dataset (necessary, but not sufficient, to preserve privacy) *(Should always be 0)*

[35] *Welcome to the SynthGauge Documentation — SynthGauge 2.1.0 documentation*. (2022). Data Science Campus. Retrieved April 26, 2024, from https://datasciencecampus.github.io/synthgauge/

[36] Raab, G., Nowok, B., & Dibben, C. (2022). Assessing, Visualizing and Improving the Utility of Synthetic Data. *Utility of Synthetic Data*. https://arxiv.org/pdf/2109.12717

Model optimisation

Tuning hyper-parameters

A hyper-parameter is like the 'setting' of a model. Decisions on hyper-parameters must be made before running the model. For example, the decision of whether or not to include an intercept term in a linear regression is like a 'hyper-parameter' for a linear regression model; as it is a decision made in the model setup which impacts its outputs.

Each of the three methods described above have different hyper-parameters. It's possible to 'tune' these hyper-parameters (select different combinations of numerical values for them) by running a model over and over again using different combinations of hyper-parameters, and seeing which performs best on predetermined metrics (for this study, these metrics are the ones for fidelity, utility and privacy described above). This process takes quite some time, so for the purposes of the feasibility study, we tuned hyper-parameters only for the most promising model for each dataset. The possible hyper-parameters for each model, and process for fine-tuning them, is described below. We outline in the results for each section which model was selected for fine-tuning, and the results before and after the tuning.

For both the GAN and MST models (which offer DP), the DP privacy budget is one parameter to be tuned. It is controlled by the privacy parameter ε, with a lower ε representing a lower risk of disclosure (a stronger privacy guarantee). The exact interpretation of the privacy parameter ε, or translation into acceptable risk thresholds, is an area of continued research and discussion. As the ONS report *Synthesising the linked 2011 census* explains, "privacy budgets are often advised to be $0.01<\epsilon<1.1$, but much larger budgets are often used in practice. For instance, the US Census Bureau used $\epsilon=19.61$[37] to create a formally private redistricting file for the 2020 Decennial Census. DP has also gained popularity in industrial settings, including at Apple, where a form of DP is used to support various iOS features. Apple uses $\epsilon=4$ to track users' favourite emojis per day..."[38].

[37] *Census Bureau Sets Key Parameters to Protect Privacy in 2020 Census Results*. (2021, June 9). Census Bureau. Retrieved April 25, 2024, from https://www.census.gov/newsroom/press-releases/2021/2020-census-key-parameters.html in Wilde et. al, 2023

[38] Wilde et. al, 2023

Conditional CART model hyperparameters

In the python implementation of the CART model for synthetic data used for this study, there is one parameter that can be tuned - the order in which variables are synthesised.[39] Variables' fidelity to the original dataset will be better preserved if they are placed earlier in the order (and so are a higher node in the Conditional CART).

We tested this by reordering variables in order of their importance for fidelity/analysis (treatment/control; outcome variables; explanatory variables; and missingness indicators), creating synthetic data, and re-benchmarking the final results.

MST model hyperparameters

For MST, there are three key parameters that can be tuned:

1. The number of iterations the model runs for to produce results: we trialled 500, 2500, and 5000.
2. The total privacy budget (epsilon): We trialled 0.01, 0.1, 1, and 10.
3. How the privacy budget is allocated: A key strength of the model is that the privacy budget can be 'allocated' to more strongly preserve some two- and three- way relationships between variables. We trialled just preserving the relationship between the treatment/control group and the first outcome variable (Pre- and Post- test outcome 1); and preserving the relationship between the treatment/control group and all outcome variables (Pre- and Post- test outcomes 1,2,3,4,5 and 6).

For each combination of parameters, we benchmarked them in terms of their mean average density scores (a measure of fidelity) and their nearest neighbour scores (a measure of privacy).

DP-GAN: As a generative adversarial neural network, the dp-gan model has a large array of hyperparameters to tune.[40] As a black box model, the exact interpretation of each hyper-parameter is not as important; instead, we empirically test the combination of numbers that produce the best results. We used a bayesian methodology to train them, through the package *optuna*.[41] Compared to a grid- or random- search approach, this increases efficiency of training by first selecting some reasonable values for hyperparameters, measuring their

---

[39] In general, Conditional CART methods can also be tuned for a wider range of parameters (for example, depth of the tree) to manage over-fitting (this was not available in the python package used for this feasibility study). If EEF decides to produce synthetic data using the Conditional CART method, they could explore the R package *synthpop* for a more customisable solution.

[40] The hyperparameters available to be tuned are: *generator_n_layers_hidden, generator_n_units_hidden, generator_nonlin, n_iter, generator_dropout, discriminator_n_layers_hidden, discriminator_n_units_hidden, discriminator_nonlin, discriminator_n_iter, discriminator_dropout:, lr, weight_decay, batch_size, encoder_max_cluster*s.

[41] Sun Microsystems. (2018). *optuna — Optuna 3.6.1 documentation*. Optuna. Retrieved May 22, 2024, from https://optuna.readthedocs.io/en/stable/reference/optuna.html

performance, then updating new values to test using the prior scores of the previous combinations of hyper-parameters.

Final evaluation

We re-benchmarked the final datasets after optimisation, as the tuned hyper-parameters should lead to improved performance (fidelity, utility and/ or privacy).

In addition to the benchmarks used initially to choose the optimal method of generating high-fidelity synthetic data (see Table 3 above), we added a t-cap privacy test. This test examines the risk that an attacker could infer the true value of a ‘target’ variable, if they had access to both the true and synthetic versions of other attributes (the ‘key’ variables). For each dataset, we selected variables that would be more likely to be available publicly for the ‘key’ variable, and a protected characteristic (where available) for the target variable.

The final benchmarks enabled us to place each dataset within the ONS’ draft framework for synthetic data (see Table 4 below)

Table 4: Proposed ONS categorisation of synthetic data[42]

Category	Sub-category	Source	Relationships Preserved	Disclosure Risk and Data Set Value
Synthetic	Structural	• Available metadata • Values derived from open sources and implausible distributions; do not match true distributions	• Data types, format	• No risk of disclosing information • Will produce worthless statistical analyses • Useful for basic testing of statistical analysis calculations (i.e., does the calculation understand the format of the data file correctly)
Synthetic	Valid	• Available metadata • Values use plausible distributions and open sources; do not match true distributions	• Data types, format, missing values • No impossible values	• Minimal risk of disclosing information • Information disclosure risk assessment should be carried out on a case-by-case basis. • Sometimes plausible ranges etc can be disclosive • Useful for advanced testing of statistical analysis calculations (e.g., does the calculation correctly process variable types and account for missing values) • Will produce worthless statistical analyses
Synthetically-Augmented	Plausible	• Real data set • Values generated from true distributions	• Data types, format, missing values • Univariate distributions	• Non-negligible risk of disclosing information • Information disclosure risk assessment should be carried out on a case-by-case basis. • Care must be taken with some kinds of data (e.g., names) • Useful for extended testing of statistical analysis calculations (e.g., does the calculation give plausible results for univariate statistics) • Will produce minimally useful statistical analyses
Synthetically-Augmented	Multivariate Plausible	• Real data set • Values generated from true distributions, preserving some relationships between them	• Data types, format, missing values • Univariate distributions • Some relationships between variables	• High risk of disclosing information • Information disclosure risk assessment should be carried out on a case-by-case basis. • Useful for teaching and testing experimental approaches to statistical analysis • Will produce somewhat useful statistical analyses
Synthetically-Augmented	Multivariate Detailed	• Real data set • Values generated from true distributions, more effort made to match joint distributions	• Data types, format, missing values • Univariate distributions • Joint distributions	• Very high risk of disclosing information • Information disclosure risk assessment should be carried out on a case-by-case basis. • Useful for teaching and testing experimental approaches to statistical analysis • Will produce somewhat useful statistical analyses

[42] Bates, A., Spakulova, I., Dove, I., & Mealor, A. (n.d.). *ONS methodology working paper series number 16 - Synthetic data pilot*. Office for National Statistics. Retrieved April 26, 2024, from https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot#synthetic-dataset-spectrum

Synthetically-Augmented	Replica	• Real data set • Values generated from true joint or conditional distributions • De-identification techniques are applied	• Data types, format, missing values • Univariate distributions • Joint distributions • Lower-level patterns	• Extremely high risk of disclosing information • Information disclosure risk assessment should be carried out on a case-by-case basis and is **critically important** • Can use in place of real data in statistical analyses; the results should be equivalent • Should be available only in secure research facilities

Results

Dataset 1: Magic Breakfast

Method selection

The cleaned Magic Breakfast dataset had 9,900 observations and 34 variables (removing unused columns and descriptors, and adding additional columns for missing variables, following the process described above).

As a reminder, the dataset used for this study was a low-fidelity version of the original Magic Breakfast results (chosen so that we could work outside the SRS, for reasons explained here). This means that existing statistical relationships had been removed from the dataset, and all continuous variables follow a standard gaussian distribution. The goal for this study was to examine to what extent we can replicate the distributions and statistical relationships present in the low-fidelity dataset; as the same methods could then replicate the relationships in the original dataset if this were available.

We generated synthetic data using the Conditional CART, MST and GAN methods, and then compared them against the benchmarks described above. Table 5 shows the results. As can be seen, Conditional CART performs best in terms of fidelity, but worst at privacy (although still within acceptable results). At the other end, GAN preserves privacy well, but has poor fidelity - this is particularly clear in the histograms, shown below and in full in the appendices. MST is in the middle; with good fidelity and privacy.

These are promising initial results, giving two feasible methods for this dataset: both Conditional CART and MST perform well enough to be useful. GAN does not, but this is unsurprising, as GAN models typically require larger datasets.

From these initial results, we chose to fine-tune the Conditional CART model to test to what extent we can improve privacy, whilst continuing to preserve fidelity. Overall, this approach provided a feasible path to produce high-fidelity synthetic data for this dataset, balancing both privacy and fidelity.

Table 5: Initial benchmarks for Magic Breakfast dataset

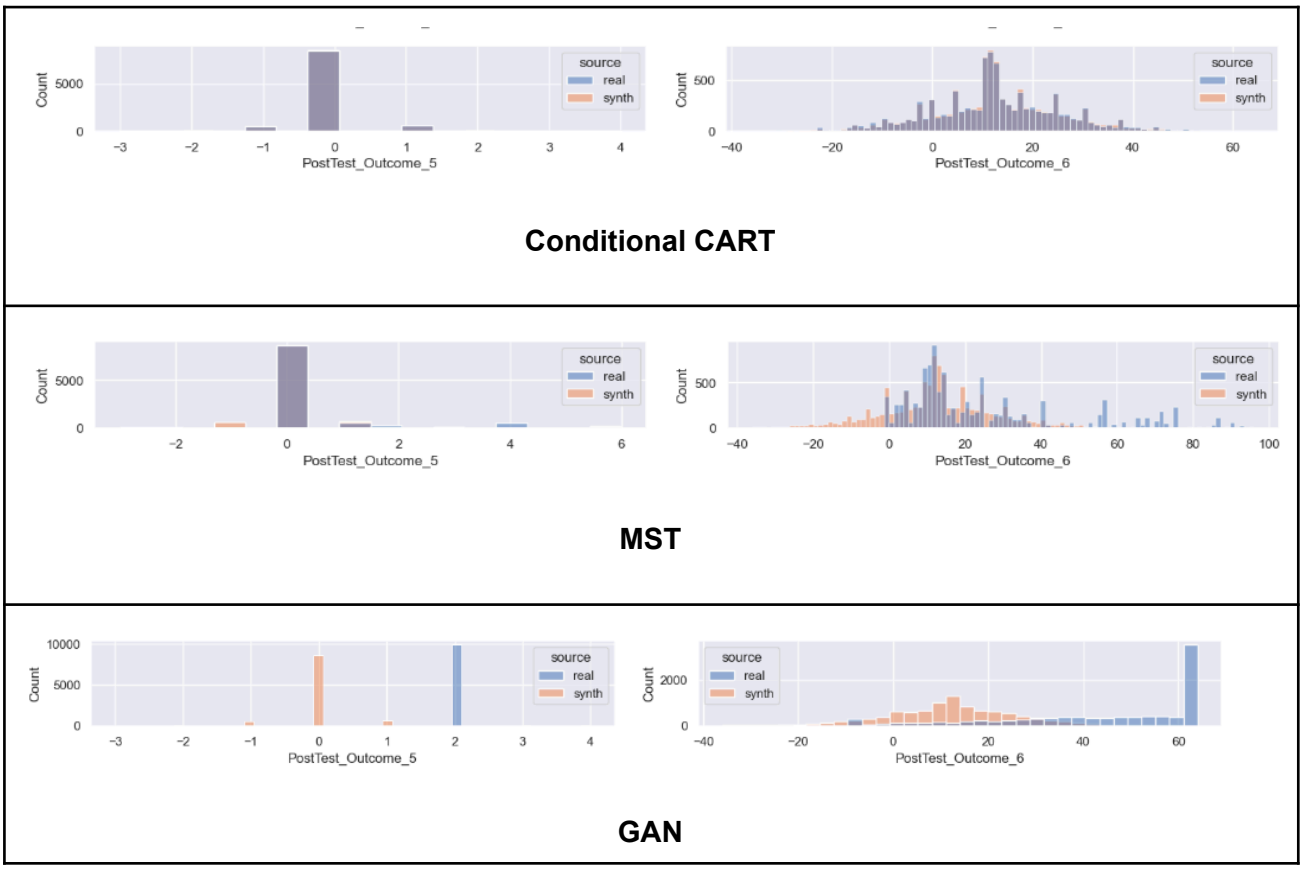
	Metric	Goal	Conditional CART	MST	GAN
Fidelity	Mean-squared difference in pairwise Pearson correlation coefficients (continuous data)	Minimise towards 0	0.00	0.00	0.00
	Correlation ratio mean-squared difference (continuous/ categorical data)	Minimise towards 0	0.00	0.00	0.00
	Mean absolute difference of feature densities	Minimise towards 0	0.03	0.05	0.82
Utility	Difference in precision on a classification task[43]	Minimise towards 0	0.00	0.00	0.10
	Difference in recall on a classification task	Minimise towards 0	0.00	0.00	0.00
	Difference in F1 on a classification task	Minimise towards 0	0.00	0.00	0.10
Privacy	Minimum Manhattan distance nearest neighbour	Maximise	1.5	6.0	24.00
	Minimum Manhattan distance nearest neighbour (outliers only)	Maximise	1.5	6.0	24.00
	Proportion of true data in synthetic data	Must be 0	0.00	0.00	0.00

In addition to the formal benchmarks, we looked at histograms of distributions and correlation matrices to get a sense of where methods may have performed better or worse. Whilst we can see that overall distributions were well preserved, MST did perform worse on continuous variables (unsurprising given the constraints of the method), where Conditional CART performed better. Illustrative examples are shown in Figure 4 below, and the full results (including correlation tables) can be seen in Appendix A.

[43] The classifier uses a support vector machine (SVM) from the *scikit learn* package, and we chose 'Post Test Outcome 1' as the variable to predict in each dataset, using all other variables. The *synthgauge* package automatically scales numeric features and one-hot encodes categorical features for the model, then runs the classifier on the synthetic and real data, and measures the difference in precision, recall and F1.

Figure 6: Illustrative histograms - Magic Breakfast dataset

Grey indicates overlap between real and synthetic data; blue/ orange represents more real/ synthetic data with that value.



Hyper-parameter tuning

As the best-performing method, we then fine-tuned the Conditional CART model to take forward as the preferred model for producing high-fidelity synthetic datasets. We did this by re-ordering the input variables to preference those most important to analyse first (as those synthesised first will have the highest fidelity):

- Treatment-Control variable ('Treatment-Control')
- Outcome variables: Those describing pre- and post- test outcomes
- Explanatory variables: First individual explanatory variables (eg FSM, Gender), and then school variables (School Context 1, 2, 3)
- Missing data variables: All variables denoting missingness in continuous variables

Whilst this did produce a very minor improvement in overall fidelity, it came at the cost of reduced privacy (see comparison in Table 6 below) - we therefore reverted to the original Conditional CART model for the final dataset. However, as the privacy loss was small, for specific purposes, this could be used to preserve important relationships.

As the Conditional CART model does not offer DP, we did not need to optimise the privacy budget.

Table 6: Comparison of pre- and post- tuned Conditional CART models (changes in bold)

	Metric	Goal	Pre-tuned Conditional CART *(final model for this study)*	Post-tuned Conditional CART
Fidelity	Mean-squared difference in pairwise Pearson correlation coefficients (continuous data)	Minimise towards 0	0.00	0.00
	Correlation ratio mean-squared difference (continuous/ categorical data)	Minimise towards 0	0.00	0.00
	Mean absolute difference of feature densities	Minimise towards 0	**0.01**	**0.00**
Utility	Difference in precision on a classification task	Minimise towards 0	0.00	0.00
	Difference in recall on a classification task	Minimise towards 0	0.00	0.00
	Difference in F1 on a classification task	Minimise towards 0	0.00	0.00
Privacy	Minimum Manhattan distance nearest neighbour	Maximise	**4.00**	**2.00**
	Minimum Manhattan distance nearest neighbour (outliers only)	Maximise	**4.00**	**2.00**
	Proportion of true data in synthetic data	Must be 0	0.00	0.00

The final test, as described above, was a T-CAP test, measuring the likelihood that an attacker could discover true values of an attribute with access to others. For this dataset, we used ‘Pupil_FSM’ as the target variable, and ‘Pupil_Gender’ and ‘PostTest_Outcome1’ as the key variables. We were measuring the likelihood that an attacker could uncover a student’s FSM status, if they had access to both the synthetic and true values of the genders of students in the study, and their year groups following the study. The TCAP score for this was 0.05 (a 5% chance the target variable could be recovered given the key variables).

Dataset 2: Inclusive learning

Method selection

The cleaned Inclusive Learning dataset had 67,200 observations and 36 variables (removing unused columns and descriptors, and adding additional columns for missing variables, following the process described above).

For this dataset, we tested both the Conditional CART and MST methods (see Table 7 for results). We attempted to train a GAN model, however, it took too long to train on this dataset - we stopped training after 48 hours, with only 24% completion. We also used a GPU to see if this improved the training time, but it did not reduce the time significantly.

As the MST model provided reasonable results, we did not invest further time in getting the GAN model to work on this dataset as part of the feasibility study. The MST model also offers DP, is easier to explain, and will be less computationally intensive than a GAN model (even if the GAN model was optimised to run more efficiently).

Table 7: Initial benchmarks for Inclusive Learning dataset

	Metric	Goal	Conditional CART	MST
Fidelity	Mean-squared difference in pairwise Pearson correlation coefficients (continuous data)	Minimise towards 0	0.00	0.00
	Correlation ratio mean-squared difference (continuous/ categorical data)	Minimise towards 0	-	-
	Mean absolute difference of feature densities	Minimise towards 0	0.00	0.02
Utility	Difference in precision on a classification task	Minimise towards 0	0.00	0.01
	Difference in recall on a classification task	Minimise towards 0	0.00	0.00
	Difference in F1 on a classification task	Minimise towards 0	0.00	0.01
Privacy	Minimum Manhattan distance nearest neighbour	Maximise	0.72	34.18
	Minimum Manhattan distance nearest neighbour (outliers only)	Maximise	0.72	34.18
	Proportion of true data in synthetic data	Must be 0	0.00	0.00

Hyper-parameter tuning

For this dataset, we tuned the MST as the final model. First, we optimised the privacy budget (by testing both the total budget, and how it was allocated). We then trialled different numbers of iterations of the model, using the best-performing setting for the privacy budget. We measured fidelity using mean average density (MAD) and privacy using n-neighbours (N-N) to understand the best parameters. As a reminder, the best set of parameters is the one which minimises MAD whilst maximising N-N. Results are shown in Table 8.

In this instance, we selected a privacy budget of 1, with all treatment/control relationships preserved, as the settings that provided the optimum mix of fidelity, utility, and privacy. However, as can be seen from the table below, all settings provided a reasonable balance, and there was not a large difference between settings.[44]

We then tested the number of iterations that the model should be trained for, using the optimal privacy budget settings. This did not make a significant difference to the outcomes observed, so we left the number of iterations to the default setting (2500).

Table 8: Tuning MST hyper-parameters

Epsilon (total privacy budget)	A single treatment/control and outcome relationship preserved	All treatment/control and outcome relationships preserved
0.01	MAD: 0.11 N-N: 832	MAD: 0.13 N-N: 825
0.1	MAD: 0.04 N-N: 801	MAD: 0.05 N-N: 807
1.0	MAD: 0.04 N-N: 798	**MAD: 0.04** **N-N: 795**
10.0	MAD: 0.04 N-N: 781	MAD: 0.04 N-N: 785

After tuning hyper-parameters, we re-benchmarked the MST model using the final dataset. Results are shown in Table 9. The T-CAP score for the final dataset was 0.08.

[44] As a reminder, a typical acceptable threshold for a privacy budget is 0.01 - 1.1; if EEF wishes to set a lower budget (and reduce the privacy risk further), these results show they could do so without a significant loss of fidelity.

Table 9: Final benchmarks for Inclusive Learning dataset

	Metric	Goal	MST (untuned)	MST - tuned (final model)
Fidelity	Mean-squared difference in pairwise Pearson correlation coefficients (continuous data)	Minimise towards 0	0.00	0.00
	Correlation ratio mean-squared difference (continuous/ categorical data)	Minimise towards 0	-	-
	Mean absolute difference of feature densities	Minimise towards 0	0.02	0.02
Utility	Difference in precision on a classification task	Minimise towards 0	0.01	0.00
	Difference in recall on a classification task	Minimise towards 0	0.00	0.00
	Difference in F1 on a classification task	Minimise towards 0	0.01	0.00
Privacy	Minimum Manhattan distance nearest neighbour	Maximise	34.18	34.15
	Minimum Manhattan distance nearest neighbour (outliers only)	Maximise	34.18	34.15
	Proportion of true data in synthetic data	Must be 0	0.00	0.00

Dataset 3: Maths Champions

Method selection

The cleaned Maths Champions dataset had 1,200 observations and 24 variables (removing unused columns and descriptors, and adding additional columns for missing variables, following the process described above). Whilst 1,200 observations is a small sample size for synthetic data generation, we wanted to test the methods, as EEF has many studies with similar sample sizes.

Table 10 shows the results of the Conditional CART and MST synthetic data generation. Due to the small sample size, we did not attempt to train a GAN model. Conditional CART and MST models struggled to produce high-fidelity synthetic data whilst preserving privacy in this dataset. Conditional CART had good fidelity, but low privacy. MST had poorer fidelity, and therefore very high privacy (as the data was not close to the original). Because initial results indicated that the models were unlikely to perform well (consistent with the general guidance that larger sample sizes are required for high-fidelity synthetic data generation), no further

action was taken for this dataset. These results show that it may only be possible to safely generate HF synthetic data for the EEF's larger datasets.

Table 10: Initial benchmarks for Maths Champions dataset

	Metric	Goal	Conditional CART	MST
Fidelity	Mean-squared difference in pairwise Pearson correlation coefficients (continuous data)	Minimise towards 0	0.00	0.00
	Correlation ratio mean-squared difference (continuous/ categorical data)	Minimise towards 0	-	-
	Mean absolute difference of feature densities	Minimise towards 0	0.1	0.05
Utility	Difference in precision on a classification task	Minimise towards 0	0.00	0.07
	Difference in recall on a classification task	Minimise towards 0	0.00	0.05
	Difference in F1 on a classification tast	Minimise towards 0	0.00	0.07
Privacy	Minimum Manhattan distance nearest neighbour	Maximise	0.64	4028
	Minimum Manhattan distance nearest neighbour (outliers only)	Maximise	0.64	4028
	Proportion of true data in synthetic data	Must be 0	0.00	0.00

Limitations

There were a number of limitations in this study. First and foremost, we used low-fidelity synthetic datasets as the basis for the study, rather than the original datasets. This was due to difficulties working in the secure environment containing the original datasets, outlined here. We believe our overall results on the feasibility of producing synthetic data are likely to be valid, as the low fidelity synthetic datasets resemble the original datasets in ways that most matter for synthetic data generation (the sample size, number of variables, and data types). However, this did present a number of limitations of the findings:

- We could not directly check for preservation of the treatment effect in the synthetic data, as this had been removed in the low fidelity data. We did, however, measure to what extent the existing statistical relationship had been preserved in the relevant variables

- Relevant groupings (particularly school groupings) had been removed. We arbitrarily created groups of observations to account for this, but similarly, cannot comment on the extent to which between- and in- group effects are maintained in the synthetic data generation process
- All continuous variables had a gaussian distribution, and there were only weak statistical relationships between variables in the low fidelity datasets. Whilst the same principles should hold in the original datasets, there is some risk that different distributions and relationships may be harder to replicate. In particular, the T-CAP tests should be re-run with the original data, as it may be easier to reconstruct variables that are highly correlated.

In addition, the study examined three datasets, out of a total of 128. Whilst we chose representative datasets on relevant characteristics, there is a chance that some datasets will behave differently to those tested as part of this feasibility study. In particular, whilst the study found that the dataset with 1,200 observations was too small to produce high fidelity synthetic data, and the dataset with 9,900 observations was sufficient, we are not able to conclude based on this study where the threshold is between these two sample sizes.

Finally, as mentioned in the methodology section, we were not able as part of the feasibility study to scan for every possible methodology to produce high-fidelity synthetic data. Our goal was to determine if there are methodologies 'good enough' to produce high fidelity synthetic data whilst preserving privacy. We leave open the possibility that there are other methodologies that would produce even better results.

Recommendations and next steps

Technical recommendations

Overall, this feasibility study shows that producing high-fidelity synthetic data is feasible on a significant proportion of EEF's datasets (those with a sample size >=9,900). Both the MST and Conditional CART models produced results that reasonably balanced fidelity and privacy, were quick to run, and relatively easy to audit and explain.

For the Magic Breakfast dataset (9,900 observations), the Conditional CART method was the best of those tested, a promising result for much of EEF's database (as the majority of projects have a similar sample size). The MST also produced reasonable results (although did not preserve fidelity as well). Using the Conditional CART method, we are able to produce 'multivariate detailed' synthetic data (using ONS draft categorisation in Table 4). We note that whilst this is high fidelity, it also carries a higher privacy risk (so the risk mitigations considered in the next section should be considered, alongside technical measures of privacy).

For the Inclusive learning dataset (sample size = 67,200), the MST method was the best of those tested. Optimising the privacy budget, we ultimately chose a value of 1, within the

range typically accepted for privacy protection.[45] Furthermore, results showed that the budget could be reduced to 0.01 without a major reduction in fidelity. Using the MST method, we are able to produce ‘multivariate detailed’ synthetic data (using ONS draft categorisation in Table 4).

There was no feasible method for producing high-fidelity synthetic data on the Maths Champions dataset (sample size = 1,200). This was somewhat expected given the small sample size, but provides a helpful approximate lower bound, under which we can empirically demonstrate that production of high-fidelity synthetic data is unlikely to be feasible.

Technically, further work will be needed to determine to what extent production of high feasibility data is feasible for trials with sample sizes between 1,200 and 9,900 observations. However, the study has demonstrated that for those datasets greater than 9,900 observations, it should be possible to produce high fidelity synthetic data using either the MST or Conditional CART method. This is consistent with guidance from the advisory board that, as a rule of thumb, high fidelity synthetic data should only be produced with datasets >5,000 observations.

Beyond the technical feasibility of producing high-fidelity synthetic data, EEF should also consider the practical requirements and any reputational or other risks associated with its production, storage and use. The section below outlines some of these risks, and we would recommend that they are incorporated into any implementation plan.

[45] Wilde et al., 2023

Implementation risks and mitigations

Production and release of high fidelity synthetic data is not merely a technical exercise. To make an informed judgement on the overall feasibility of producing and releasing high fidelity synthetic datasets for its database of RCTs, EEF should consider both actual and perceived risks.

Actual risk: privacy breach

The key actual risk to consider is an individual's data ending up in the public domain through publication of high-feasibility datasets. There are two main ways this could happen:

- **Accidental release of data:** Production of high fidelity synthetic data is a complex exercise, requiring expertise and judgement to maintain a good balance of fidelity and privacy. Getting this wrong could result in accidental release of data.
- **Bad actor reconstruction attack:** With advances in technology, true data can be reconstructed with increasingly little information. The amount of information required for reconstruction is likely to continue to decrease with advances in the field.

Even with correctly produced high fidelity data, it should be noted that the methods described here (including DP) reduce, but do not eliminate, the risk of data leakage in the datasets - EEF will need to accept some risk (however small) in production. To manage these risks, there are a number of actions to take, beyond the technical privacy measures of DP and benchmarking discussed above:

1. **Production and assurance:** All high fidelity synthetic datasets should be produced by a data scientist (or equivalent), with an appropriate quality assurance process in place (including review from a colleague/ third party). We estimate that it took 1 FTE day of a data scientist's time per dataset, once the methodology had been finalised, and this should be built into EEF's resourcing going forward. In addition, EEF should plan to regularly review their existing processes in light of developments in the field, to ensure that existing datasets and metrics are resistant to the latest reconstruction attacks.
2. **Control access to data and models:**
    a. EEF should consider restrictions on who can access high fidelity data (for example, by continuing to store the datasets in the IDS, the UKDS; or releasing only to accredited researchers with clear usage agreements in place outside the SRS). Options for restricting access are explored in more detail below. This is a key measure to reduce the potential consequence of a data leak, and minimise the risk of reconstruction attack.
    b. To reduce the risks of reconstruction attack, synthetic data generation models should be kept within a secure environment (eg, IDS, UKDS), at the same level of security as the original data.
    c. To reduce the risk of reconstruction attack, only ranges for fidelity, utility and privacy metrics should be published, rather than exact values.

Perceived risk: privacy breach

Even where there is no actual privacy breach, high-fidelity synthetic data is not a commonly understood concept. Without appropriate communications, there is a risk that EEF's RCT study participants fear a privacy breach has occurred, and their data has been released in public, or has been used for a purpose that they did not originally agree to.

EEF should consider potential risks to its reputation and viability of its studies when considering production of high fidelity synthetic data. Whilst EEF will have the best view of how to manage these relationships and communications, we recommend that EEF engage with prior participants in their studies before releasing any high fidelity synthetic datasets, to ensure they understand what the datasets are, why they are being produced, and how their privacy is protected.

Perceived risk: inaccurate data

Even with high fidelity synthetic data, there will be some statistical differences between synthetic and original data. If synthetic data is used for research without appropriate disclaimers or controls, researchers may come to different conclusions than if they use the original data. This could reduce trust in EEF's evidence base, as the research outcome may vary. To reduce this risk, EEF could consider:

1. Requiring anyone using high fidelity synthetic datasets to undergo a short training on how to use the data appropriately, prior to gaining access.
2. Restricting use of high fidelity synthetic datasets for pre-publication analysis only, to refine users' code and speed up their final analysis. Any publications would still require access to the original data, minimising the risk of public loss of trust in the dataset.

Box 5: Synthetic data in the US Census Bureau[46]

In 2020, the US Census Bureau announced its intention to start using DP as a method of disclosure avoidance in detailed census tables. The decision, and misunderstandings surrounding its reasons and implications, resulted in ongoing controversy (including legal action).

From the perspective of those working on the census, the decision was a relatively minor one: there had been a process of adding noise to aid with disclosure control to the Census since 1990, and DP represented a more sophisticated way of doing this. Furthermore, the Bureau went out of its way to be consultative; announcing their intent early and seeking views from a range of stakeholders.

However, as Boy & Sarathy note, “Many data users did not know what to make of this information or the request for feedback. They were uncertain of what the Census Bureau was asking for, let alone why. A broad swath of stakeholders expressed disbelief that the bureau intended to alter the data it collected; many of these stakeholders appeared unaware that the bureau had been intentionally altering the data for disclosure avoidance purposes since 1990.”[47]

The controversy is explored in detail in the paper *Differential Perspectives: Epistemic Disconnect Surrounding the US Census Bureau’s Use of Differential Privacy.* A key takeaway is the need for careful communication around synthetic data and differential privacy, that takes into account the starting assumptions of the audience about the true data, and their capacity to absorb technical information about the proposed change. It highlights the risks - to trust in the data, and to the reputation of the organisation - that can arise if implementation goes poorly.

[46] Boyd, D., & Sarathi, J. (2022, April 20). *Differential Perspectives: Epistemic Disconnects Surrounding the US Census Bureau’s Use of Differential Privacy*. SSRN. Retrieved May 14, 2024, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4077426

[47] Boyd & Sarathi, 2022, p3

Implementation recommendations

If EEF chooses to publish high fidelity synthetic datasets, they will have to answer two questions - who can access the data, and what can they do with it? Options and recommendations are discussed below.

Who can access the data?

Option 1: The complete high fidelity synthetic datasets are released publicly (eg, on EEF website) - *Not recommended*

This is the option that provides the greatest access to datasets, and has previously been used in the UK (see, for example the Simulacrum synthetic dataset, of ~2 million cancer patient records held by the NHS).[48] However, in these instances the datasets have been much larger than typical EEF datasets, with correspondingly lower privacy risk. As a high risk option, this will also require more resources to maintain (keeping up to date with emerging risks, and assisting with public enquiries for the datasets).

Option 2: The high fidelity synthetic datasets are released outside of a secure environment like IDS, but only to accredited researchers - *Recommended for further investigation*

This option, and option 3, provide the best balance between enabling access and managing risk. This is the process followed by the Scottish Longitudinal Study - researchers must still complete the SRS training to get accredited, but can then have immediate access to high fidelity datasets.[49] This would enable the EEF to put in place some risk mitigation (training and usage agreements, for example), but still enable researchers to access datasets much more quickly than within the secure setting.

This could be combined with the public release of low-fidelity datasets, or subsets of high-fidelity datasets, to enable researchers to gauge the utility of the full dataset prior to undergoing training.

Option 3: The high fidelity synthetic datasets are stored within the IDS, but at a lower security setting than the original datasets - *Recommended for further investigation*

EEF could explore producing high-fidelity datasets to be stored within the IDS, but at a lower security level than the original datasets. Depending on IDS access policies, this could mean

[48] *Available data*. (n.d.). simulacrum.healthdatainsight.org.uk. Retrieved May 14, 2024, from https://simulacrum.healthdatainsight.org.uk/available-data/

[49] How to access SLS synthetic data. (2017, November 9). Scottish Longitudinal Study Development & Support Unit. Retrieved April 26, 2024, from https://sls.lscs.ac.uk/guides-resources/synthetic-data/how-to-access-sls-synthetic-data/

researchers are able to access the datasets more quickly than the original data, whilst retaining the benefits of a safe setting. However, further work is required (likely when the IDS is more developed) to understand any benefits to researchers of improved access.

Option 4: The high fidelity synthetic datasets are stored within the IDS, at the same security setting as the original data - *Not recommended*

EEF could produce and store synthetic datasets with the original data. This is clearly the lowest risk option. However, it presents no additional benefits to researchers (as it will take just as long to access the original datasets, which can then be used for publication immediately). We therefore do not recommend this for further enquiry, as it would take resources from the EEF without providing a tangible benefit to researchers.

What can researchers do with the data?

As discussed above, EEF will also need to decide what researchers can do with the data. Whilst the overall fidelity in this study for the medium and large datasets were strong, some statistical difference remained (particularly at the variable level). To maintain trust in their RCT results, EEF could consider limiting usage - for example, for pre - publication analysis only, or for exploring only specific statistical relationships.

References

(n.d.). synthpop - Generate synthetic data. Retrieved April 25, 2024, from https://synthpop.org.uk/

Available data. (n.d.). simulacrum.healthdatainsight.org.uk. Retrieved May 14, 2024, from https://simulacrum.healthdatainsight.org.uk/available-data/

Bates, A., Spakulova, I., Dove, I., & Mealor, A. (n.d.). *ONS methodology working paper series number 16 - Synthetic data pilot*. Office for National Statistics. Retrieved April 26, 2024, from https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot#synthetic-dataset-spectrum

Bowen, C., & Snoke, J. (2020). Comparative Study of Differentially Private Synthetic Data Algorithms from the NIST PSCR Differential Privacy Synthetic Data Challenge. *arXiv*. arXiv:1911.12704v3.

Boyd, D., & Sarathi, J. (2022, April 20). *Differential Perspectives: Epistemic Disconnects Surrounding the US Census Bureau's Use of Differential Privacy*. SSRN. Retrieved May 14, 2024, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4077426

Calcraft, P., Thomas, I., Maglicic, M., & Sutherland, A. (2021, December 14). *Accelerating public policy research with synthetic data*. Behavioural Insights Team | ADR. https://www.adruk.org/fileadmin/uploads/adruk/Documents/Accelerating_public_policy_research_with_synthetic_data_December_2021.pdf

Census Bureau Sets Key Parameters to Protect Privacy in 2020 Census Results. (2021, June 9). Census Bureau. Retrieved April 25, 2024, from

https://www.census.gov/newsroom/press-releases/2021/2020-census-key-parameters.html

Freie Universitat Berlin. (n.d.). *Random Forest*. Blogs@FU-Berlin. Retrieved April 25, 2024, from https://blogs.fu-berlin.de/reseda/random-forest

Hazy Ltd. (2017, November 9). *synthpop README*. Github: hazy/synthpop. Retrieved April 25, 2024, from https://github.com/hazy/synthpop?tab=readme-ov-file#readme

How to access SLS synthetic data. (2017, November 9). Scottish Longitudinal Study Development & Support Unit. Retrieved April 26, 2024, from https://sls.lscs.ac.uk/guides-resources/synthetic-data/how-to-access-sls-synthetic-data/

Jarrett, D., Cebere, B., Liu, T., & Curth, A. (2022). *HyperImpute: Generalized Iterative Imputation with Automatic Model Selection*. https://arxiv.org/abs/2206.07769

Jordan, J., Houssiau, F., Cherubin, F., Cohen, S., Szpruch, L., Bottarelli, M., & C. Maple. (2022). *Synthetic Data - what, why and how?* arxgiv>Computer Science> Machine Learning. https://arxiv.org/abs/2205.03257

McKenna, R., Miklau, G., & Sheldon, D. (2021). *Winning the NIST contest: a scalable and general approach to differentially private synthetic data*. http://arxiv.org/abs/2108.04978

Nowok, B., Raab, G., & Dibben, C. (2017). Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R. *Statistical Journal of the IAOS*, *33*, 785-796. ResearchGate. 10.3233/SJI-150153

Pathare, A., Mangrulkar, R., Suvarna, K., Parekh, A., Thakur, G., & Gawade, A. (2023). Comparison of tabular synthetic data generation techniques using propensity and cluster log metric. *International Journal of Information Management Data Insights*, *100177*(3), 1-12. Elsevier. 10.1016/j.jjimei.2023.100177

Raab, G. (2022). Utility and Disclosure Risk for Differentially Private Synthetic Categorical Data. *Privacy in Statistical Databases*. https://arxiv.org/pdf/2206.01362

Raab, G., Nowok, B., & Dibben, C. (2022). Assessing, Visualizing and Improving the Utility of Synthetic Data. *Utility of Synthetic Data*. https://arxiv.org/pdf/2109.12717

Sarathy, J., & Boyd, D. (2022, March 15). Differential Perspectives: Epistemic Disconnects Surrounding the US Census Bureau's Use of Differential Privacy. *Harvard Data Science Review*.

Sun Microsystems. (2018). *optuna — Optuna 3.6.1 documentation*. Optuna. Retrieved May 22, 2024, from https://optuna.readthedocs.io/en/stable/reference/optuna.html

Synthetic data for public good | Data Science Campus. (2019, February 21). Data Science Campus. Retrieved April 24, 2024, from https://datasciencecampus.ons.gov.uk/projects/synthetic-data-for-public-good/

Synthetic LS data. (2017, November 9). CALLS -HUB. Retrieved April 26, 2024, from https://calls.ac.uk/guides-resources/synthetic-ls-data/

UK Office for National Statistics. (2024, 04 24). *About the Integrated Data Service*. Available Tooling. https://integrateddataservice.gov.uk/available-tooling

Welcome to the SynthGauge Documentation — SynthGauge 2.1.0 documentation. (2022). Data Science Campus. Retrieved April 26, 2024, from https://datasciencecampus.github.io/synthgauge/

Wilde, H., Lawrence, M., Houssiau, F., & Daniel, O. (2023). *Synthesising the linked 2011 Census and deaths dataset while preserving its confidentiality*. Data Science for the Public Good. https://datasciencecampus.ons.gov.uk/synthesising-the-linked-2011-census-and-deaths-dataset-while-preserving-its-confidentiality/

Xiong, L., Haoran Li, & Jiang, X. (2014). *Differentially Private Synthesization of Mult-Dimensional Data using Copula Functions*.

https://www.researchgate.net/publication/268793256_Differentially_Private_Synthesization_of_Multi-Dimensional_Data_using_Copula_Functions

Appendix A: Data cleaning

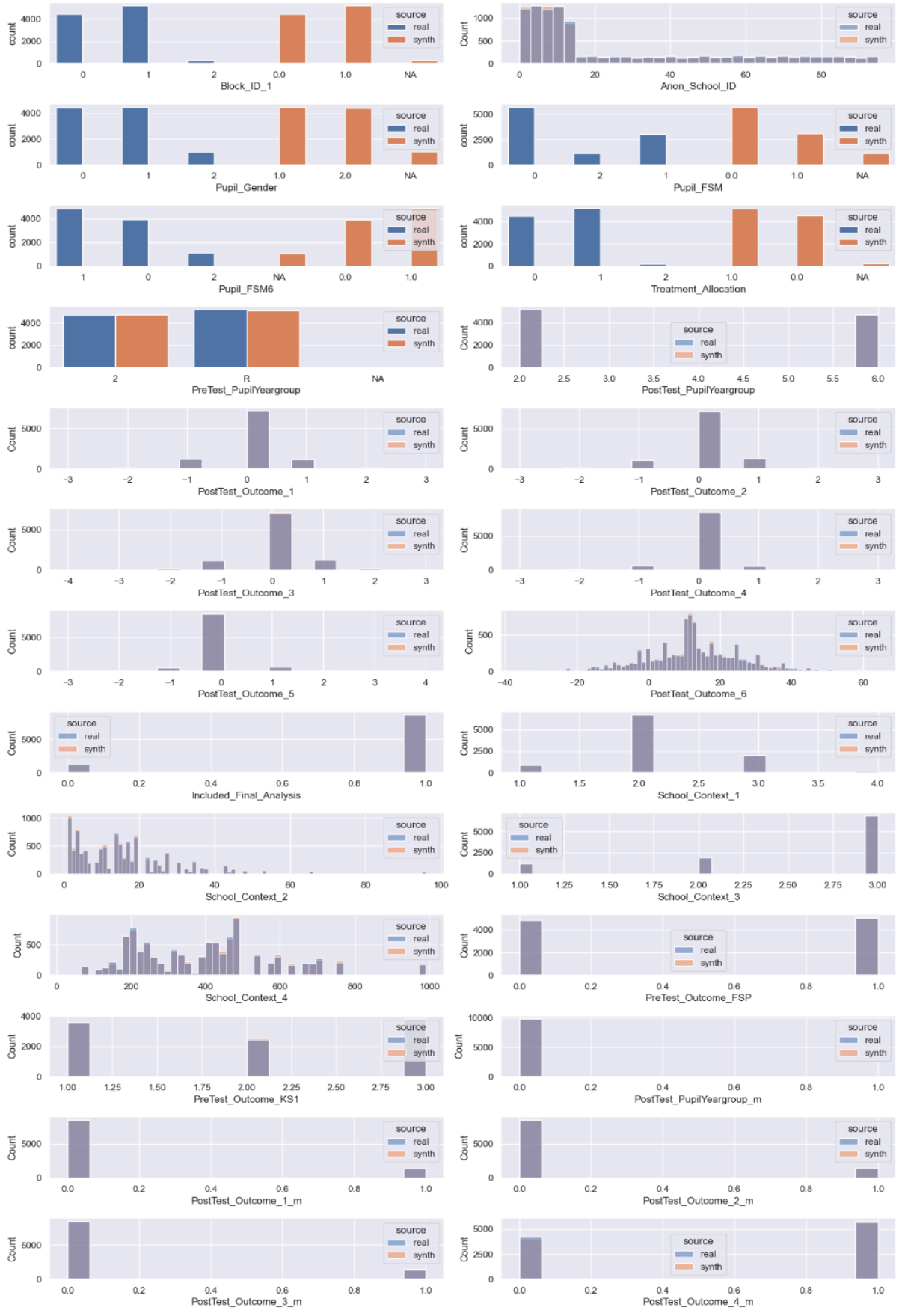
Remove redundant data:

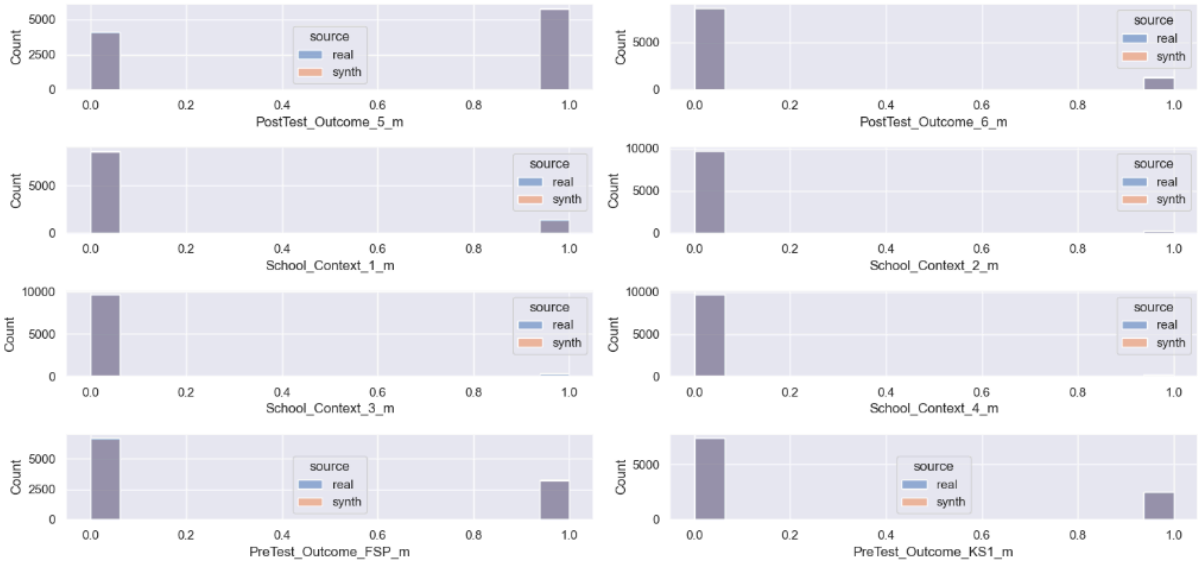
    - Remove columns that apply to the project as a whole (eg project name, project type)
    - Remove unused columns
    - Remove columns wholly dependent on others: Several columns were text descriptors of other columns. We removed the text descriptors, as the relationship can be learnt from one column, and the descriptors added back in following synthetic data generation.
    - Remove unit IDs: We removed columns that were a unique ID for the row, as generators automatically treat rows as the unit of analysis. We retained IDs that served to cluster rows (eg, school identifiers)
- Manage data types
    - Ensuring variables are labelled as the correct data type (integer, float, object), and removing text fields and errors
    - Ensuring correct numeric/ non-numeric markers: Where numbers had been used to indicate a categorical variable, we changed these to non-numeric markers, so that the algorithms do not learn the order between markers. Similarly, we converted ordinal variable to numeric markers, so that order is preserved in synthetic data generation.
- Remove one-hot encoding: Some outcomes had been one-hot encoded. These were converted back into an ordinal series, as this makes it easier to preserve relationships during data generation
- Impute missing values: The packages we used required all missing values to be imputed prior to synthetic data generation. Because we wanted to retain missingness as a feature in the synthetic dataset, we:
    - Created a new category for categorical variables, replacing missing data with an “NA” category
    - Created a binary indicator for continuous variables, creating new columns that indicate the proportion of ‘missingness’, and imputing values in the original column to retain statistical relationships. For imputation we used the imputer *hyperimpute*. It uses an iterative model of classification (for categorical variables) and regression (for continuous variables) to impute missing values, such that the overall statistical properties of each column are preserved as closely as possible.[50]
    - We then trained a model and generated synthetic data for both the column with the continuous data, and the column containing the proportion of missingness.
    - Finally, we merged the synthetic columns (replacing the generated values with missing values, where missing values had been generated in the binary column), to create a synthetic dataset containing a representative proportion of missing values.

[50] (Jarrett et al., 2022)

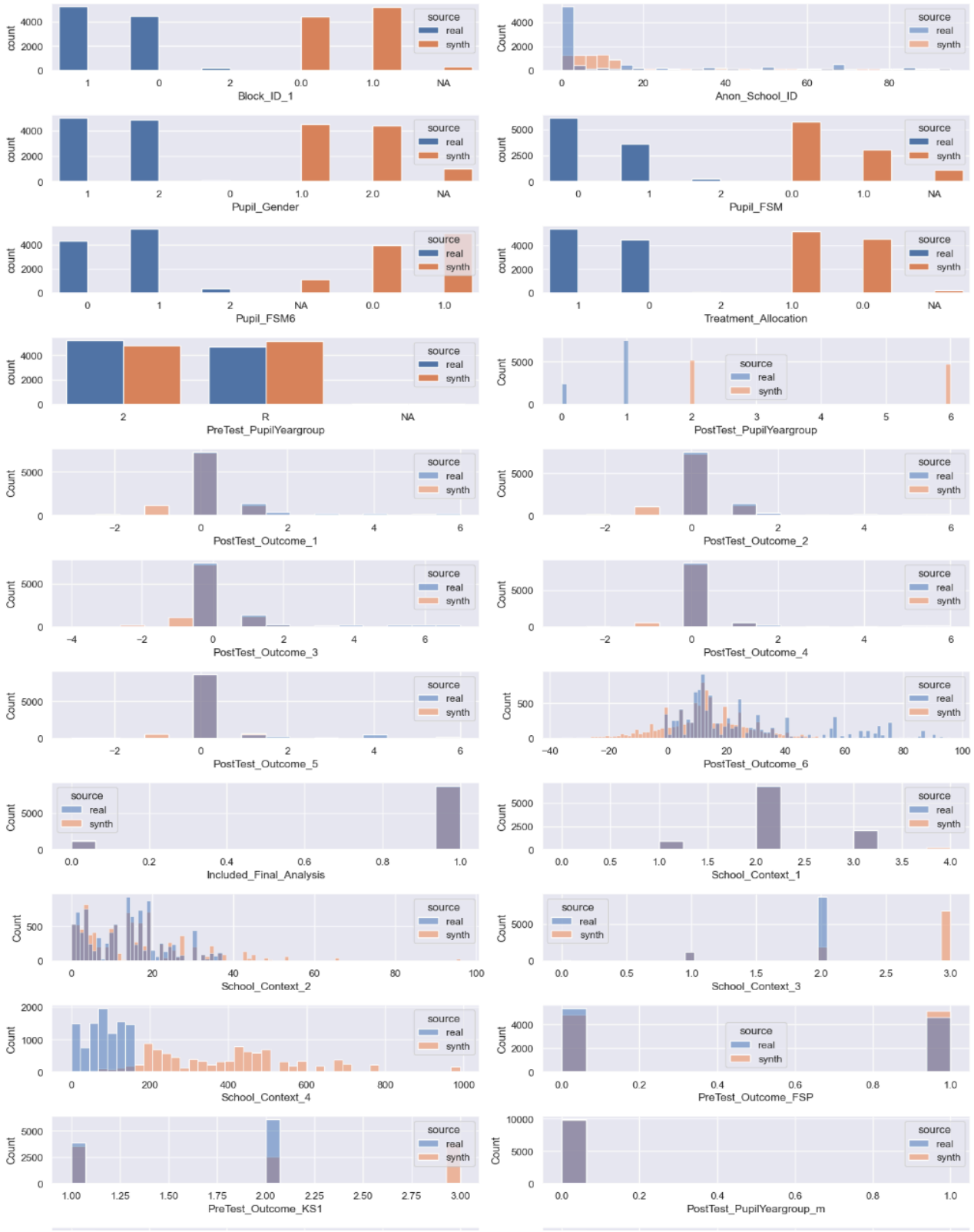
Appendix B: Detailed Results

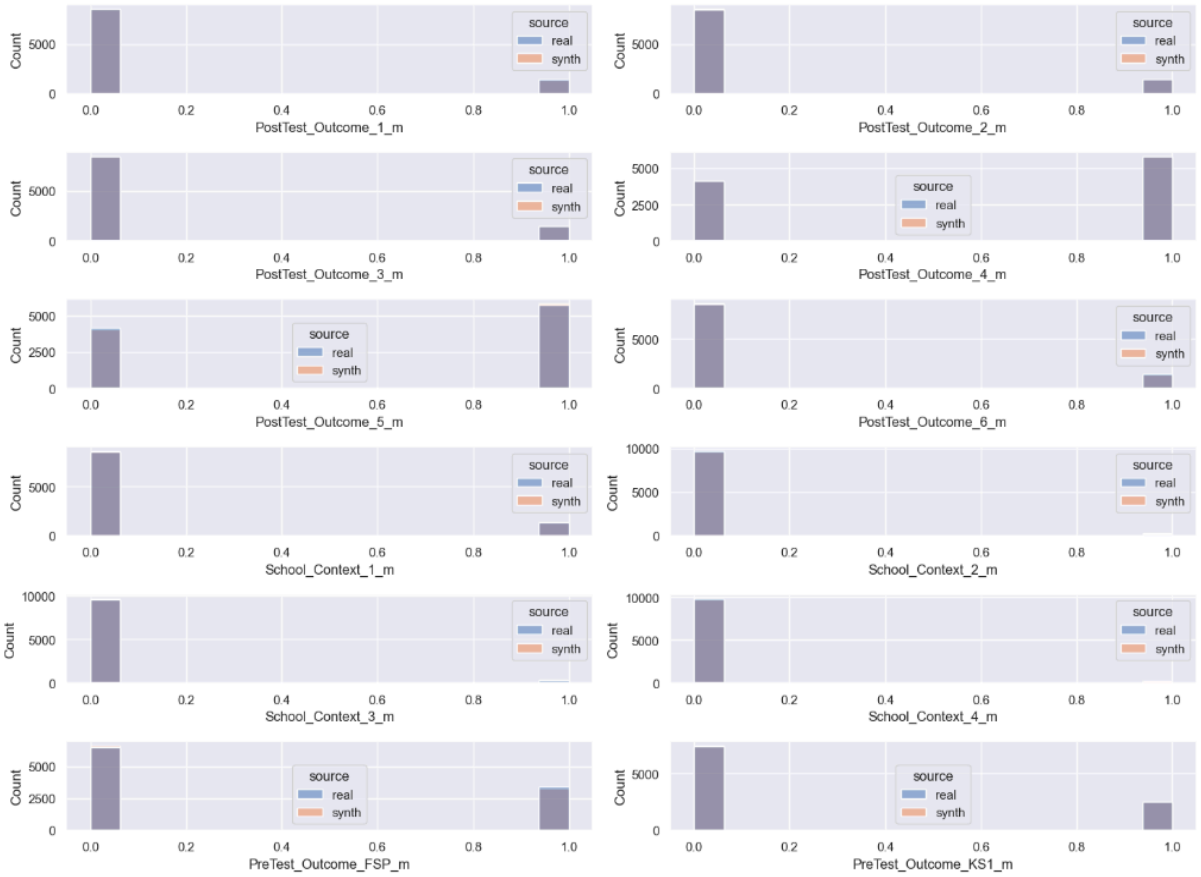
Magic Breakfast - Conditional CART method - Histogram - Untuned



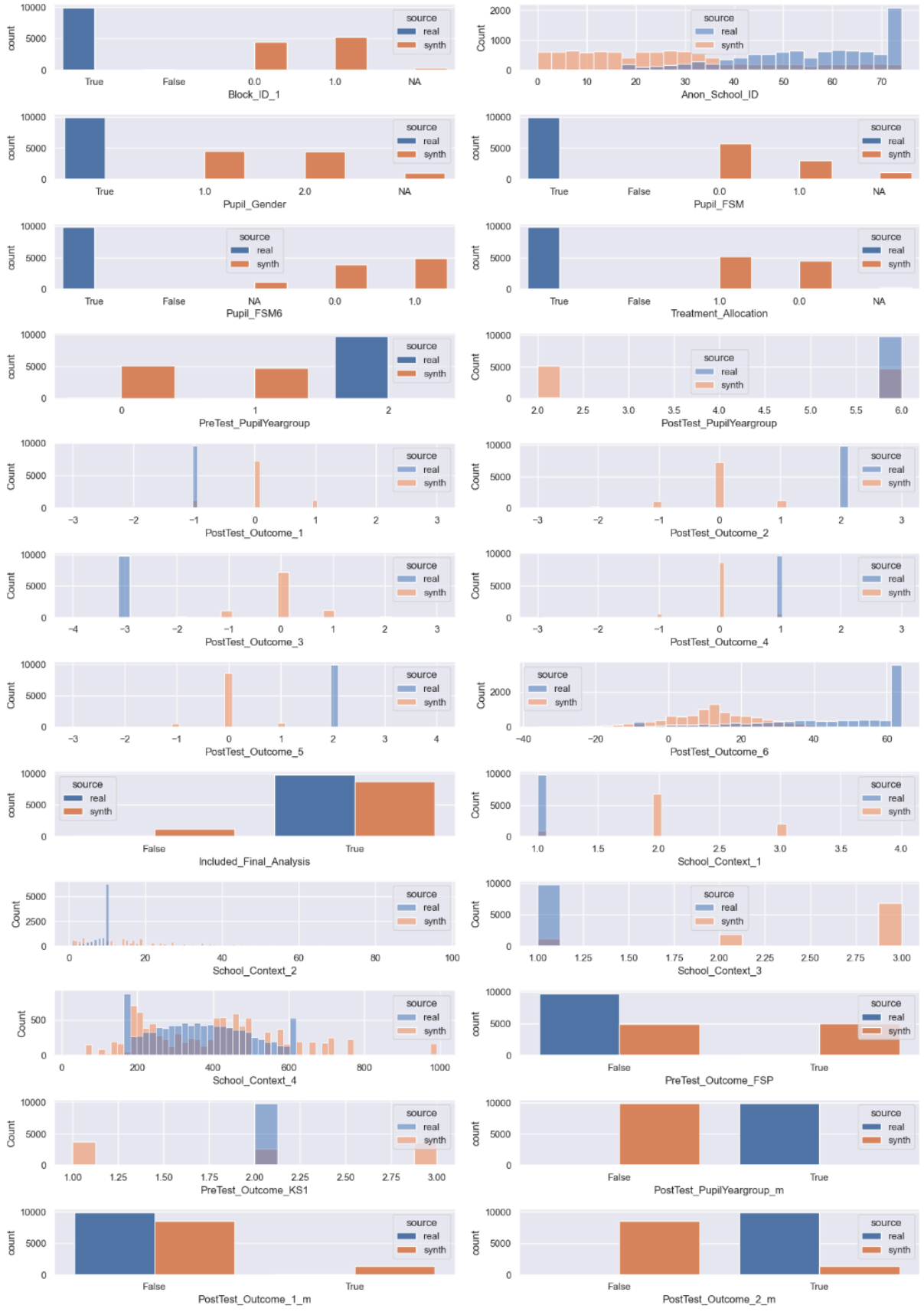

Count
source
real
synth
PostTest_Outcome_5_m
PostTest_Outcome_6_m
School_Context_1_m
School_Context_2_m
School_Context_3_m
School_Context_4_m
PreTest_Outcome_FSP_m
PreTest_Outcome_KS1_m

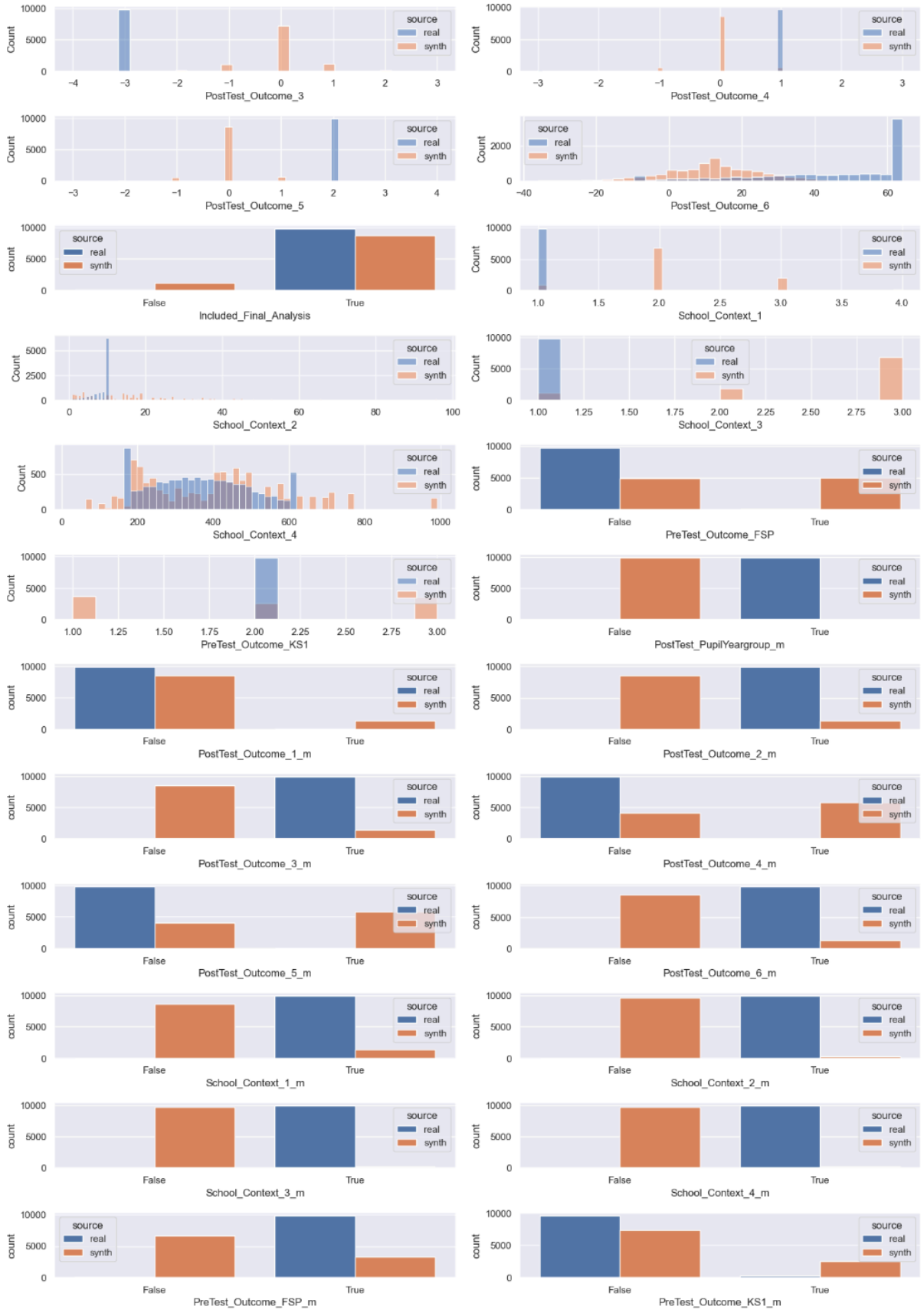
Magic Breakfast - MST method - Histogram - Untuned



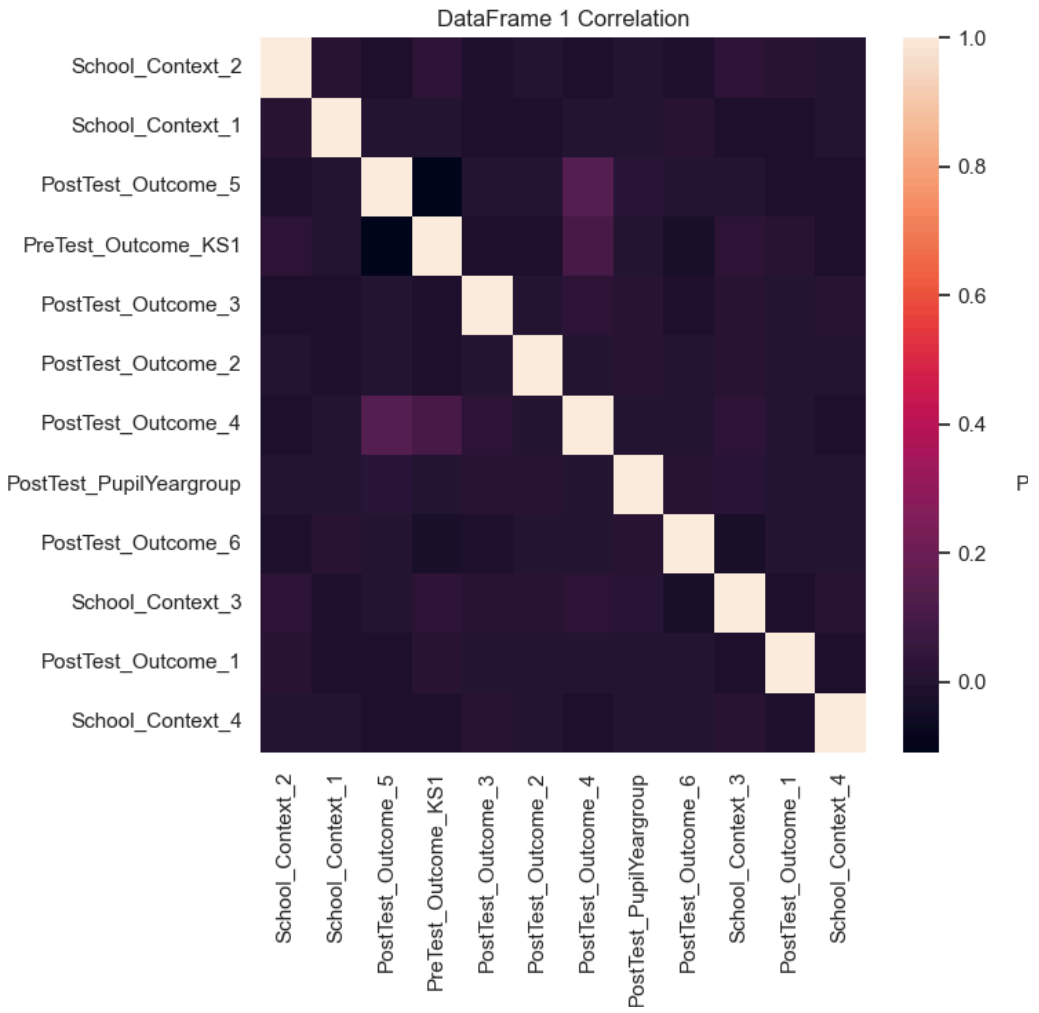

Count
source
real
synth
PostTest_Outcome_1_m
PostTest_Outcome_2_m
PostTest_Outcome_3_m
PostTest_Outcome_4_m
PostTest_Outcome_5_m
PostTest_Outcome_6_m
School_Context_1_m
School_Context_2_m
School_Context_3_m
School_Context_4_m
PreTest_Outcome_FSP_m
PreTest_Outcome_KS1_m

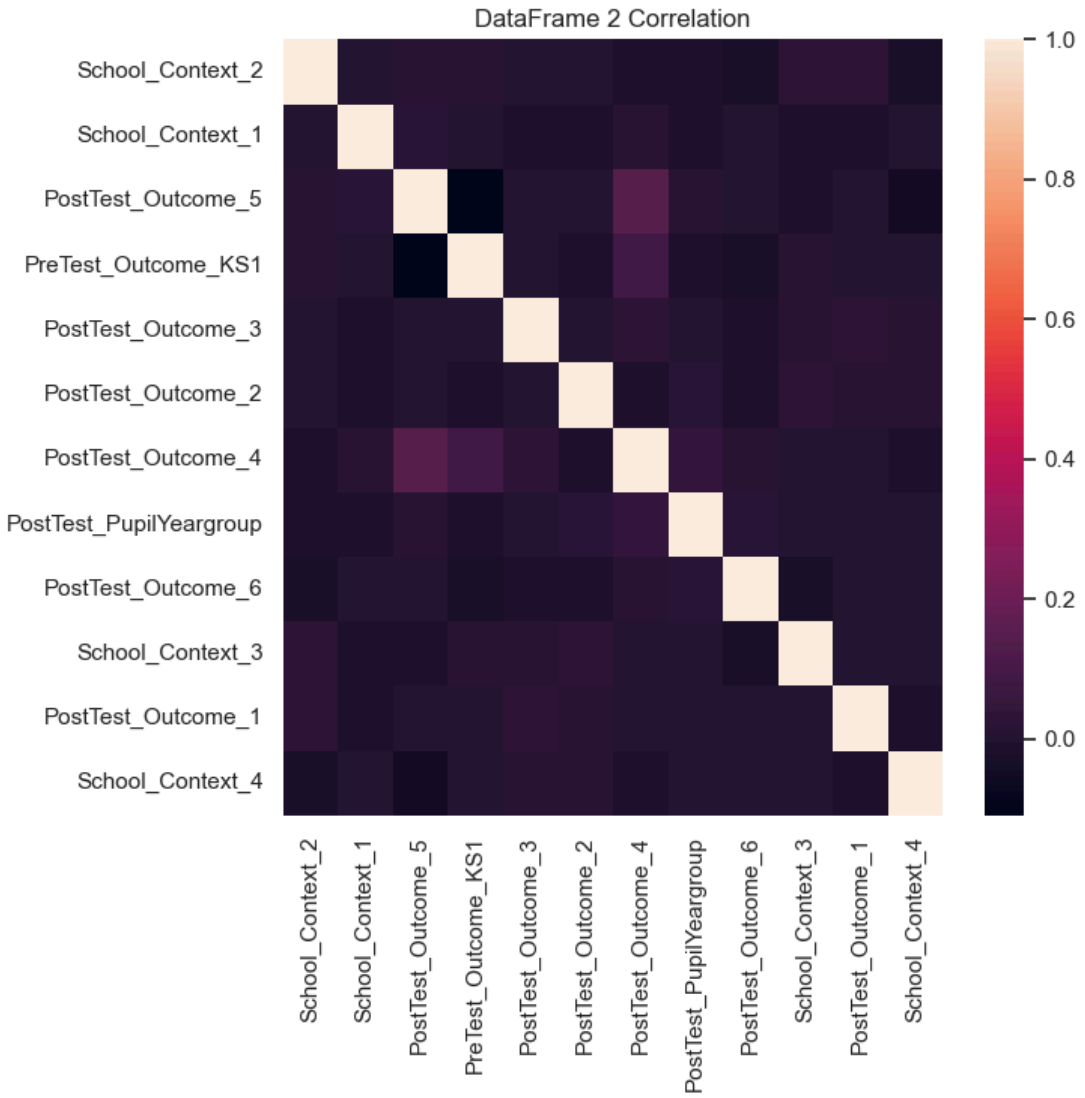
Magic Breakfast - GAN method - Histogram - Untuned

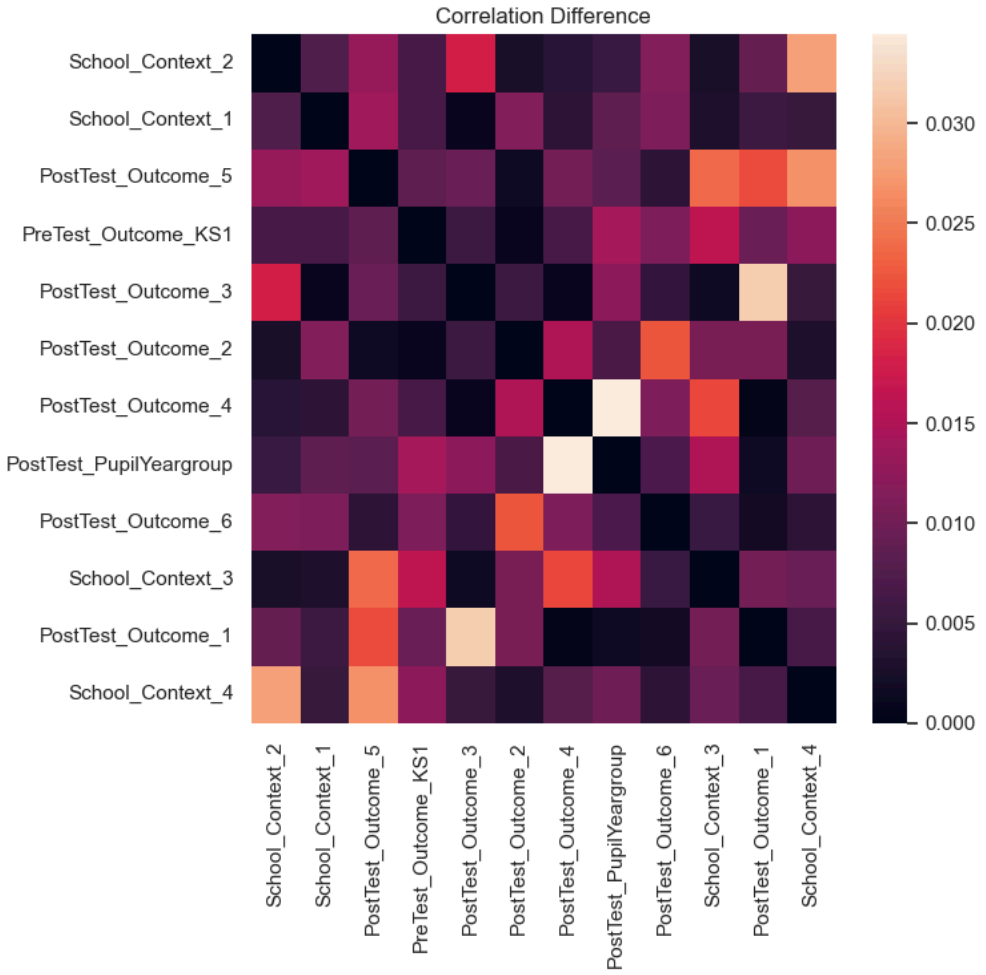



PostTest_Outcome_3
PostTest_Outcome_4
PostTest_Outcome_5
PostTest_Outcome_6
Included_Final_Analysis
School_Context_1
School_Context_2
School_Context_3
School_Context_4
PreTest_Outcome_FSP
PreTest_Outcome_KS1
PostTest_PupilYeargroup_m
PostTest_Outcome_1_m
PostTest_Outcome_2_m
PostTest_Outcome_3_m
PostTest_Outcome_4_m
PostTest_Outcome_5_m
PostTest_Outcome_6_m
School_Context_1_m
School_Context_2_m
School_Context_3_m
School_Context_4_m
PreTest_Outcome_FSP_m
PreTest_Outcome_KS1_m
source
real
synth
Count
count
False
True

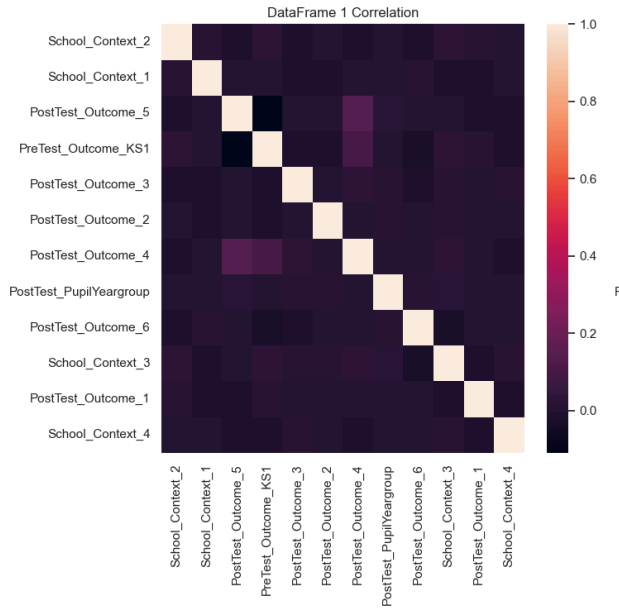
Magic Breakfast - Correlation matrices - Conditional CART untuned

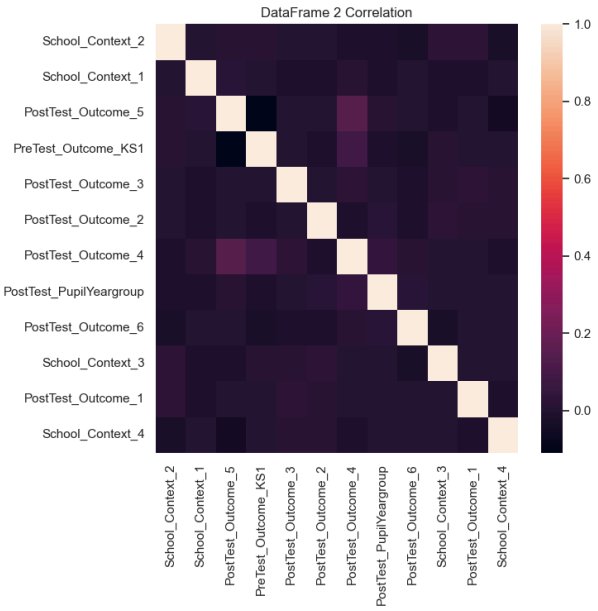


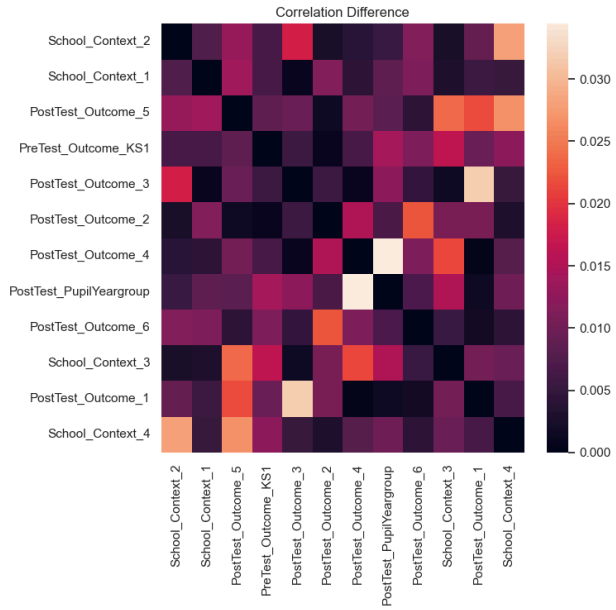

DataFrame 2 Correlation
School_Context_2
School_Context_1
PostTest_Outcome_5
PreTest_Outcome_KS1
PostTest_Outcome_3
PostTest_Outcome_2
PostTest_Outcome_4
PostTest_PupilYeargroup
PostTest_Outcome_6
School_Context_3
PostTest_Outcome_1
School_Context_4
1.0
0.8
0.6
0.4
0.2
0.0


Correlation Difference
School_Context_2
School_Context_1
PostTest_Outcome_5
PreTest_Outcome_KS1
PostTest_Outcome_3
PostTest_Outcome_2
PostTest_Outcome_4
PostTest_PupilYeargroup
PostTest_Outcome_6
School_Context_3
PostTest_Outcome_1
School_Context_4
0.030
0.025
0.020
0.015
0.010
0.005
0.000

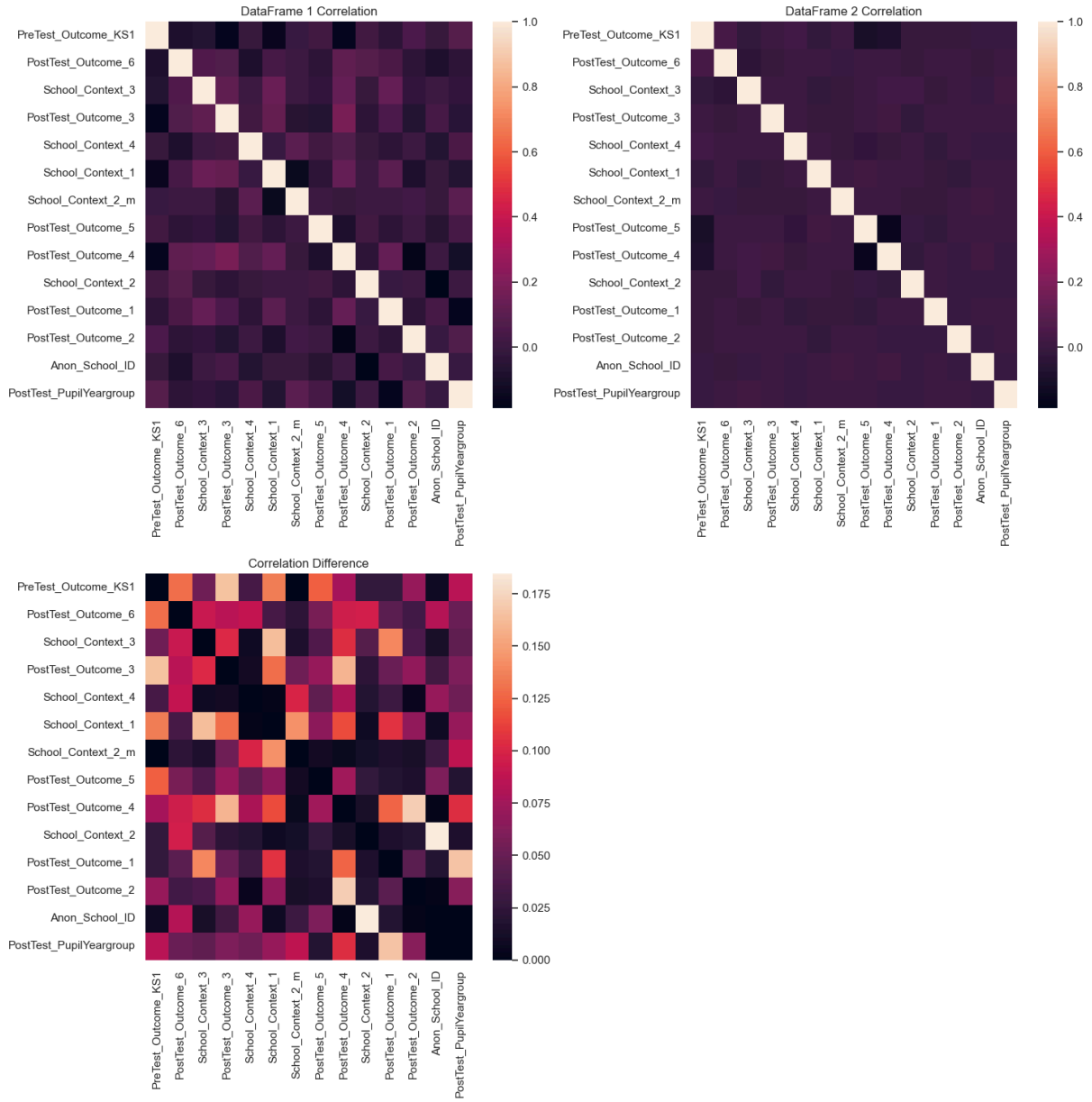
Magic Breakfast - Correlation Difference - MST - Untuned



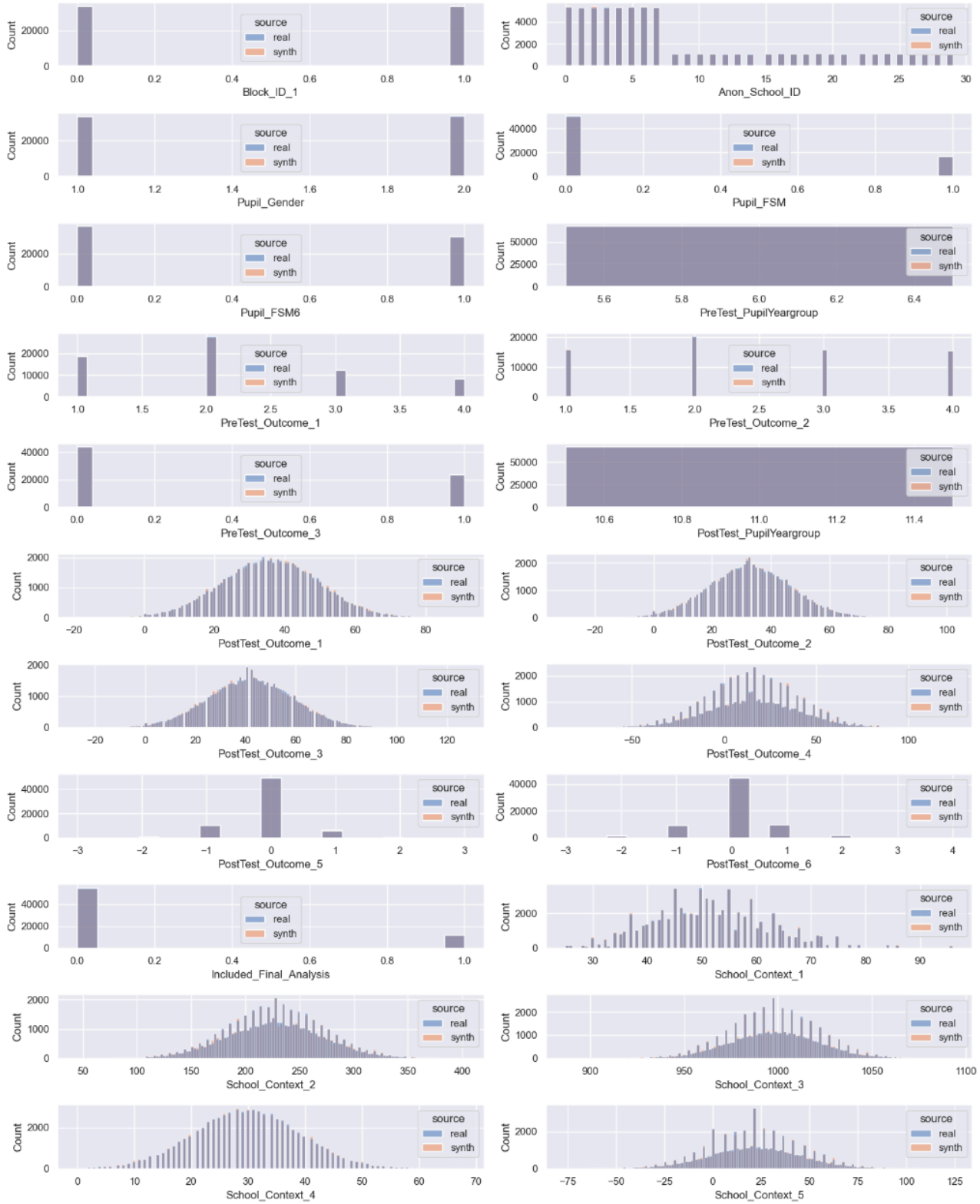




Magic Breakfast - Correlation Difference - GAN method - Untuned

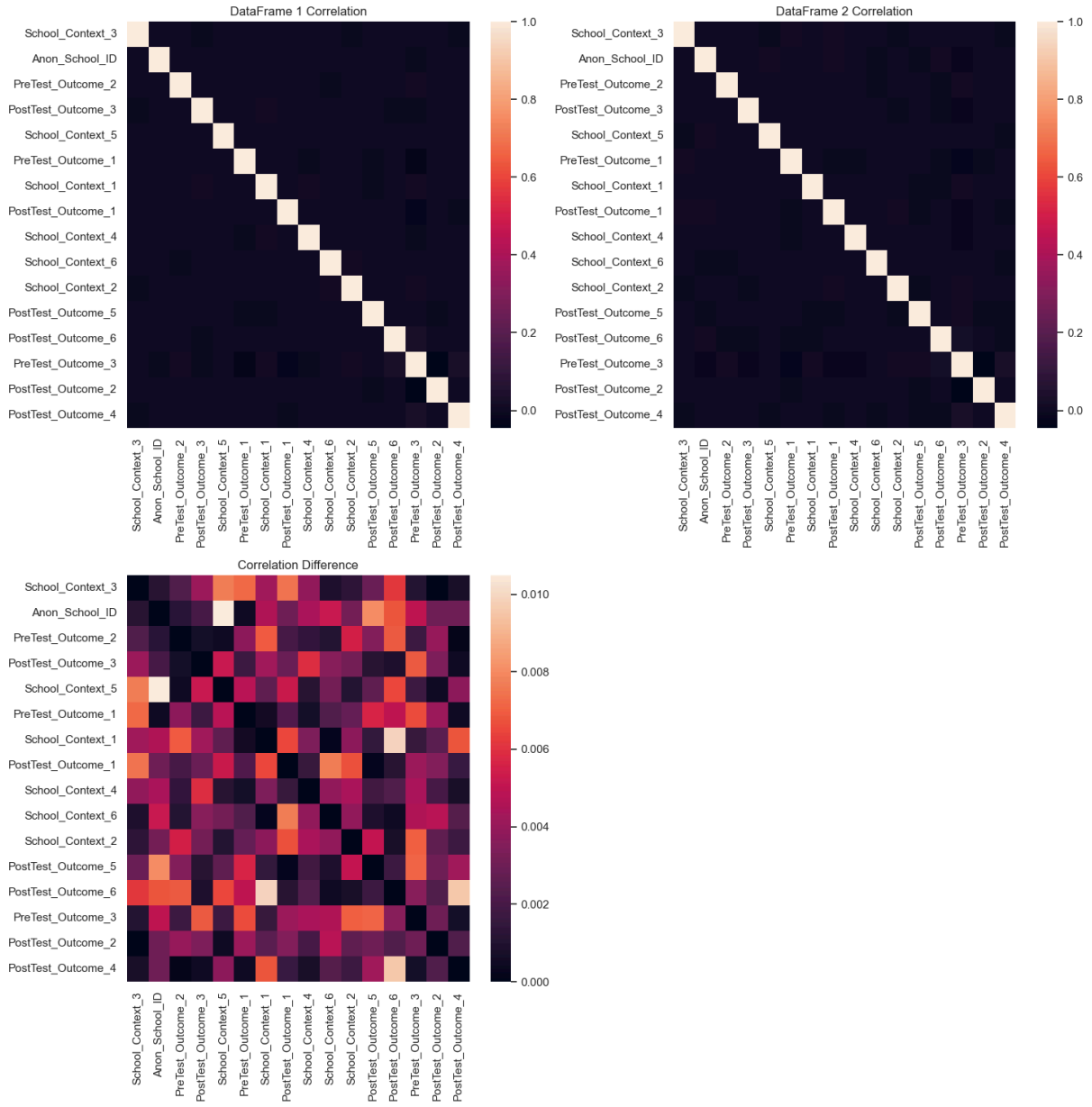


Inclusive Learning - Conditional CART model - Untuned - Histogram

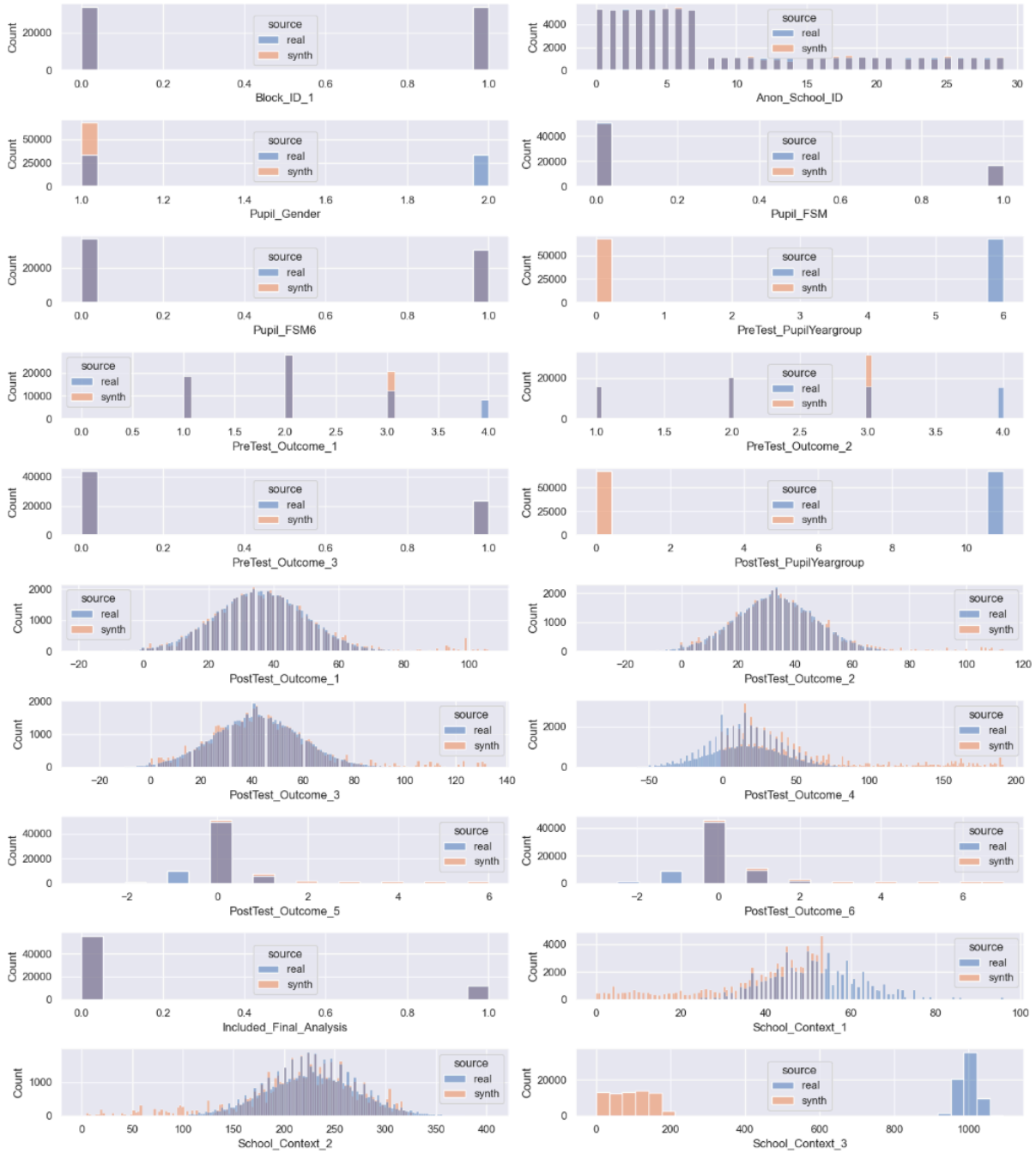



Count
source
real
synth
School_Context_6
Pupil_FSM_m
Pupil_FSM6_m
PreTest_Outcome_1_m
PreTest_Outcome_2_m
PreTest_Outcome_3_m
PostTest_Outcome_2_m
PostTest_Outcome_3_m
PostTest_Outcome_4_m
PostTest_Outcome_6_m
School_Context_1_m
School_Context_2_m
School_Context_3_m
School_Context_5_m

Inclusive Learning - Conditional CART model - Untuned - Correlation plots

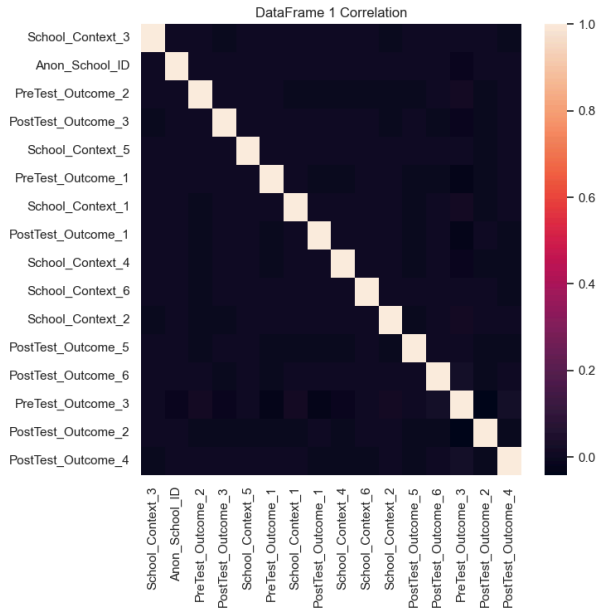


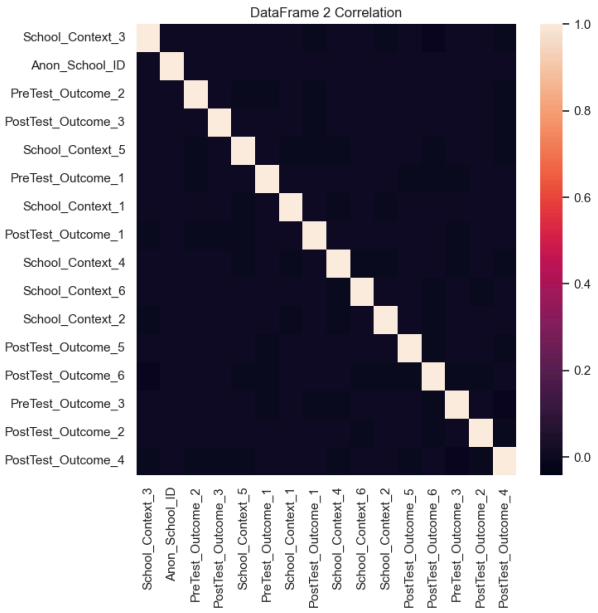
Inclusive Learning - MST model - Untuned - Histogram

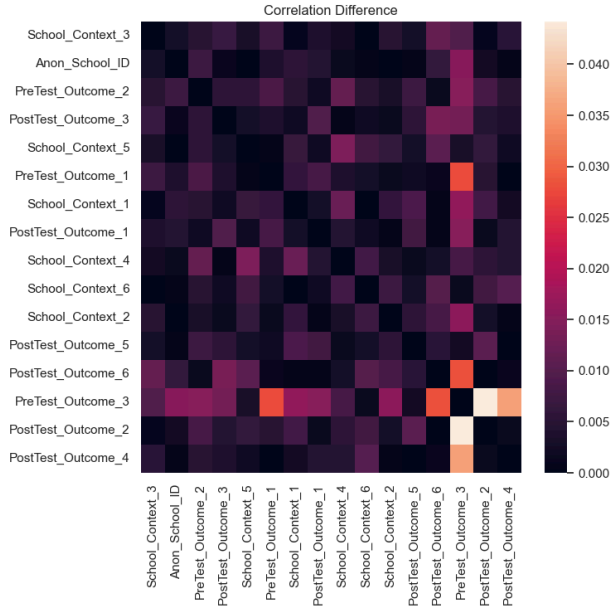



source
real
synth
Count
School_Context_4
School_Context_5
School_Context_6
Pupil_FSM_m
Pupil_FSM6_m
PreTest_Outcome_1_m
PreTest_Outcome_2_m
PreTest_Outcome_3_m
PostTest_Outcome_2_m
PostTest_Outcome_3_m
PostTest_Outcome_4_m
PostTest_Outcome_6_m
School_Context_1_m
School_Context_2_m
School_Context_3_m
School_Context_5_m

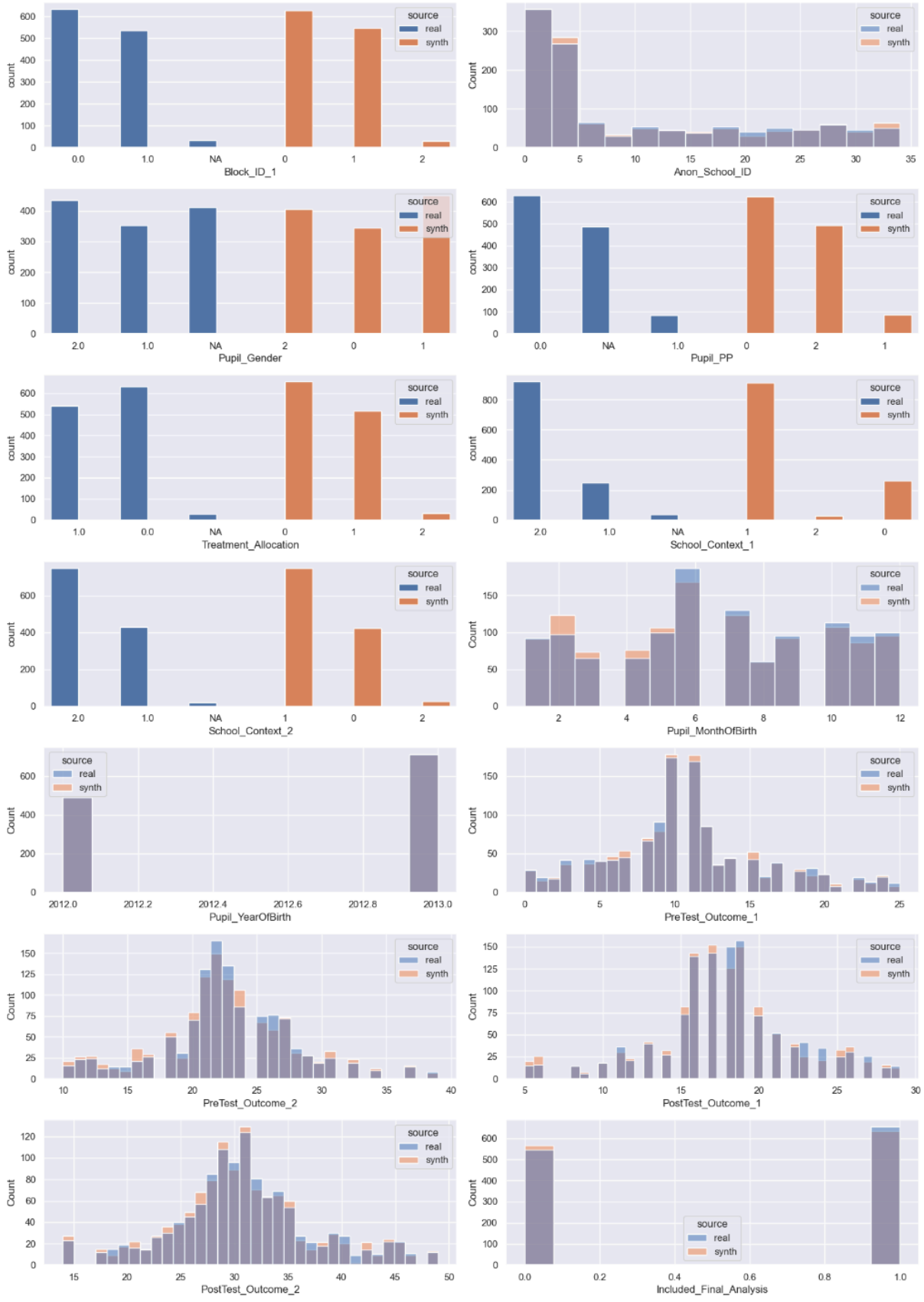
Inclusive Learning - MST model - Untuned - Correlation plots

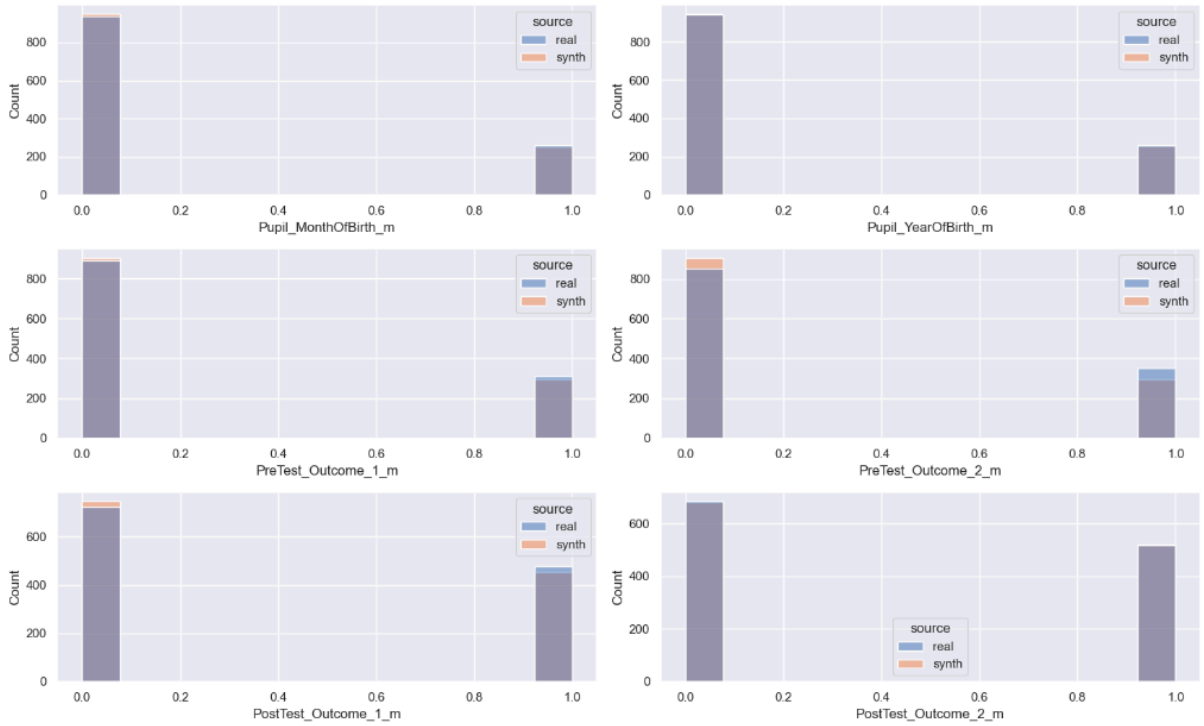




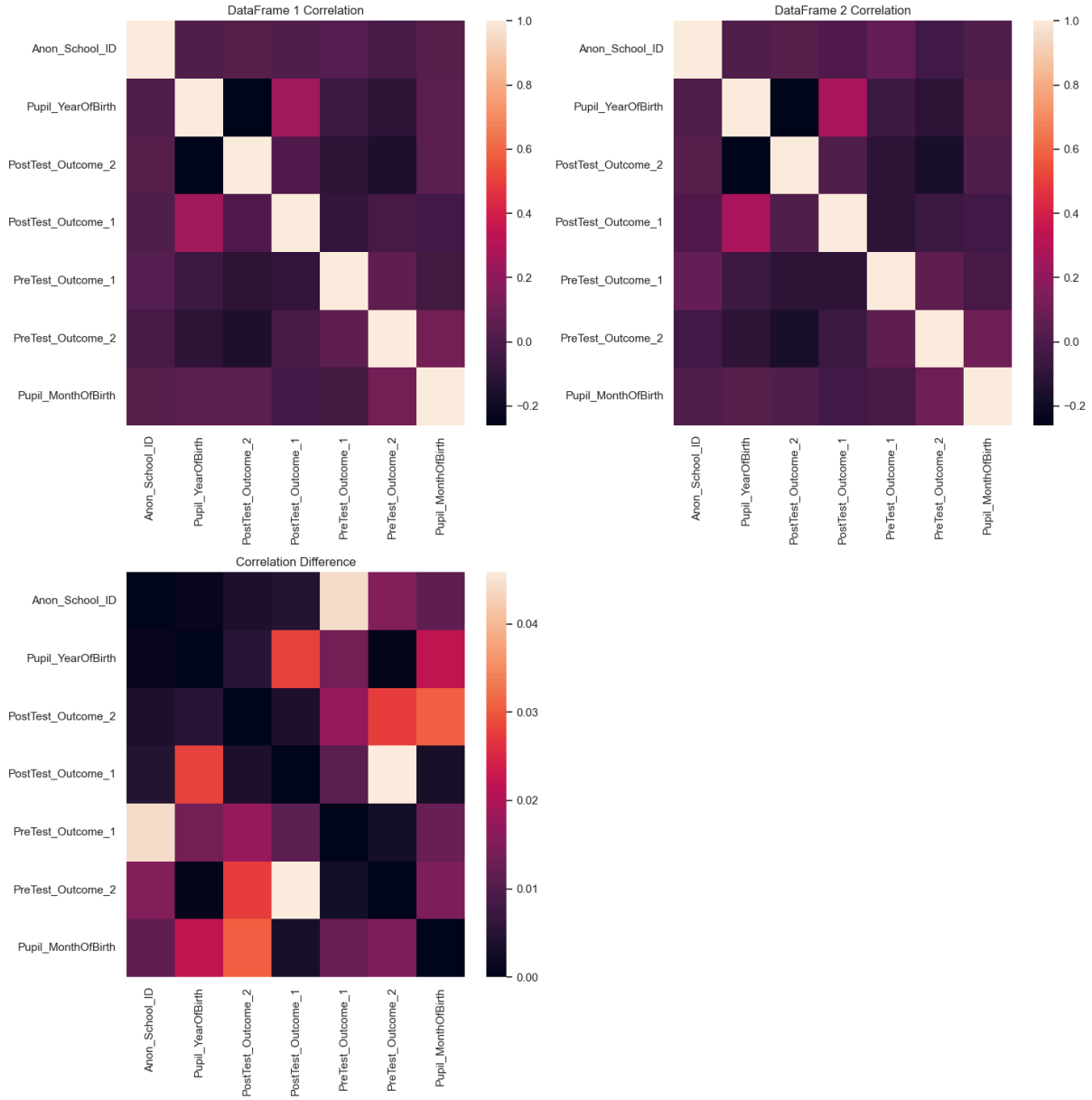


Maths Champions - Conditional CART model - Histogram

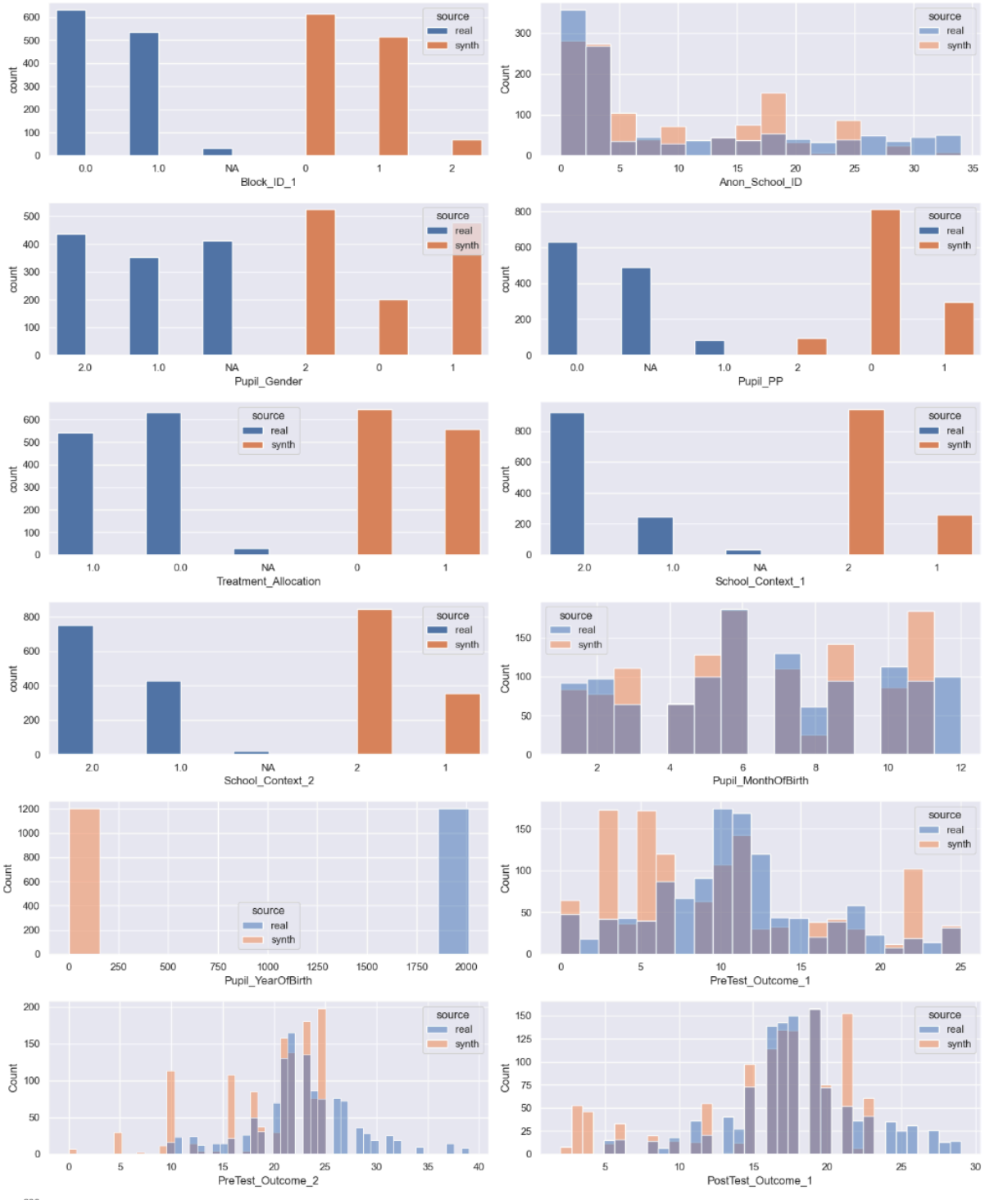


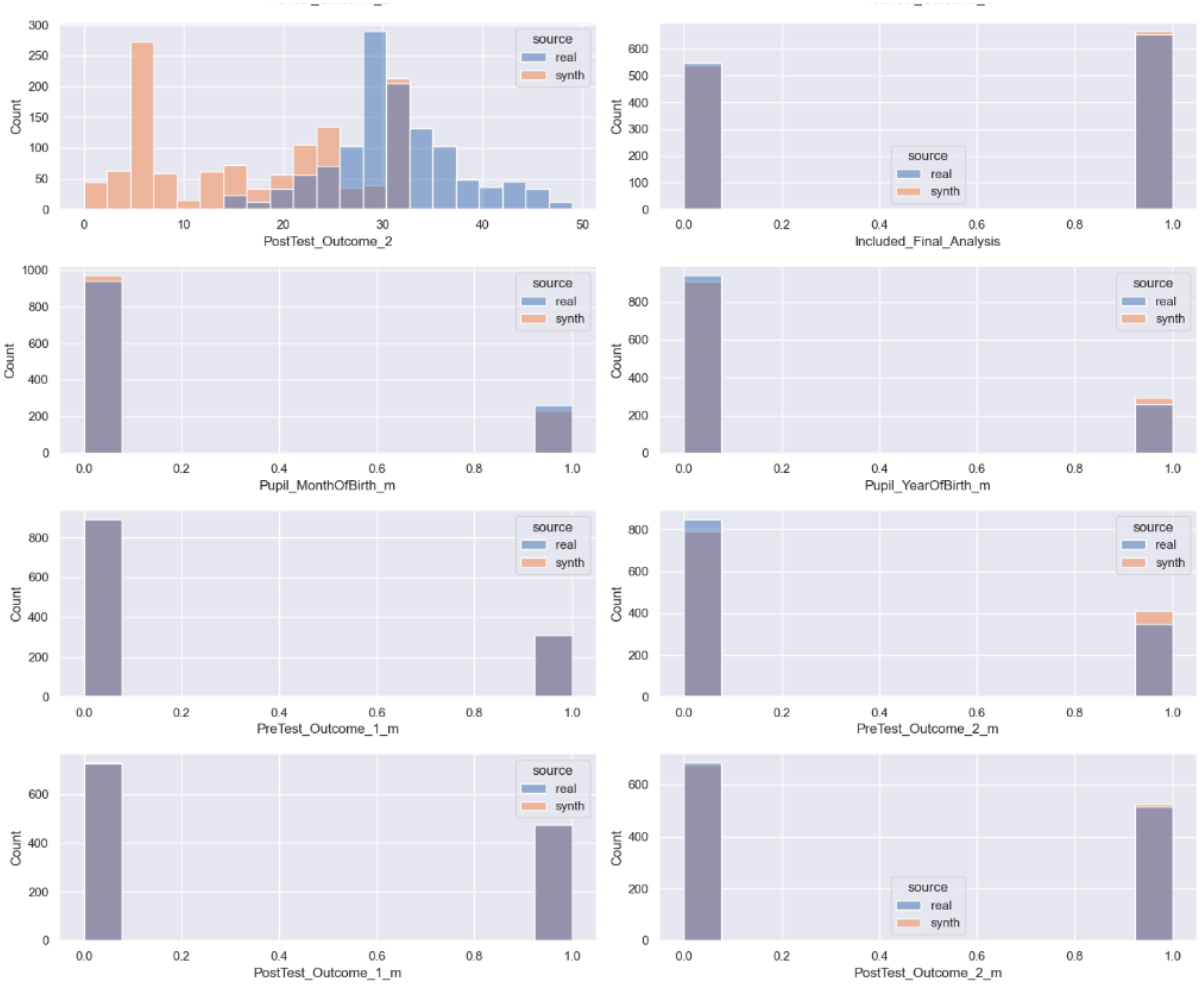

source
real
synth
Count
Pupil_MonthOfBirth_m
Pupil_YearOfBirth_m
PreTest_Outcome_1_m
PreTest_Outcome_2_m
PostTest_Outcome_1_m
PostTest_Outcome_2_m

Maths Champions - Conditional CART model - Correlation



Maths Champions - MST model - Histogram




source
real
synth
Count
PostTest_Outcome_2
Included_Final_Analysis
Pupil_MonthOfBirth_m
Pupil_YearOfBirth_m
PreTest_Outcome_1_m
PreTest_Outcome_2_m
PostTest_Outcome_1_m
PostTest_Outcome_2_m

Maths Champions - MST model - Correlation