(1) Evaluation of Synthetic instances:

Think about the results and for each of the four use cases in the guide, write whether you think that synthesis technique on that data, with those utility and privacy results, would produce a suitable result for that use case.

We used the methods SMOTE ,PCA ,AutoEncoders, COPULA and EFPA_HIST to generate the synthetic data for both satgpa.csv(toy data set) and ACP survey data(real data set) .
Evaluated the utility metrics and privacy metrics of the generated synthetic data with all these methods. After interpreting the results (see page 3 for full details of methods, metrics and results of both the data sets), the ability to use fully private methods (COPULA and EFPA_HIST) and partially private methods (SMOTE, PCA, AE) for the use cases mentioned in the guide varies in terms of utility and confidentiality.

Method categories	*Use Case1 (Releasing synthetic microdata to the public)*	*Use Case2 (Testing Analysis)*	*Use Case3 (Education)*	*Use Case4 (Testing Technology)*	*Specific tools used to generate data*	*Evaluation Methods used*	
						Utility	*Privacy*
SMOTE	Not suitable for this use case, as SMOTE produces High utility and medium Confidentiality	Not suitable for this use case, as SMOTE produces High utility and medium Confidentiality	Yes, SMOTE is suitable for this uses ,as SMOTE produces High utility and medium Confidentiality	Yes, SMOTE is suitable for this uses ,as SMOTE produces High utility and medium Confidentiality (might be a bit advanced as it has high utility).	We developed our own simulation pipeline in "R" programming	Overall summary metrics, pMSE, SPECKS, Ratio pMSE Pearson Pairwise Correlation Coefficients and Random cuts	Full intersections, Pairwise Intersections, Mean of Fuzzy Distacne, Mean of Inverse Fuzzy Distance
PCA	Not suitable for this use case, as PCA produces High utility and medium Confidentiality	Not suitable for this use case, as PCA produces High utility and medium Confidentiality	Yes, PCA suitable for this use case, as PCA produces High utility and medium Confidentiality	Yes, SMOTE is suitable for this uses ,as SMOTE produces High utility and medium Confidentiality (might be a bit advanced as it has high utility).	We developed our own simulation pipeline in "R" programming	Same as above	Same as above
AE (Auto Encoders)	Not suitable for this use case, as AE produces High utility and medium Confidentiality	Not suitable for this use case, as AE produces High utility and medium Confidentiality	Yes, AE suitable for this use case, as AE produces High utility and medium Confidentiality	Yes, AE suitable for this use case, as AE produces High utility and medium Confidentiality	We devloped a simulation pipeline in "R" programming and also we used H2o package for this	Same as above	Same as above
EFPA_HIST	Not suitable for this use case, as EFPA_HIST produces medium utility and medium Confidentiality	Not suitable for this use case, as EFPA_HIST produces medium utility and medium Confidentiality	Not suitable for this use case, as EFPA_HIST produces medium utility and medium Confidentiality	Yes EFPA_HIST is suitable for this use case, as EFPA_HIST produces medium utility and high Confidentiality	We developed our own simulation pipeline in "R" programming	Same as above	Same as above
COPOLA	Not suitable for this use case, as COPULA produces medium utility and medium Confidentiality	Not suitable for this use case, as COPULA produces medium utility and medium Confidentiality	Not suitable for this use case, as COPULA produces medium utility and medium Confidentiality	Yes COPULA is suitable for this use case, as COPULA produces medium utility and high Confidentiality	We developed our own simulation pipeline in "R" programming	Same as above	Same as above

1. ***Creating both a fully synthetic and a partially synthetic file***

We generated the fully synthetic data for both satgpa data set and ACS data set with all the features.

We also generated the partially synthetic data for ACS data set by taking sample of the data set and synthesized few features ("SEX","MARST","RACE", "AGE","INCTOT","INCWAGE","INCEARN")

2. Demonstrating evidence of tuning (adjusting model parameters to improve performance)

Used the parameter combinations for generating synthetic data for the given data sets but not optimised the parameters to generate the synthetic data.

	Original Data	Fully Synthetic	Partially synthetic	Parameter Tuning	Methods Used
1	satgpa Data set	Yes(we fully synthesized this data set)	No	Run the multiple iterations with different combinations of the parameters	SMOTE,PCA,AE,COPULA, HIST
2	ACS Data set	Yes(we fully synthesized this data set)	Yes(for a sample of the ACS data, we partially synthesized for some sensitve variables (3 categorical and 3 numerical) from all variables)	Run with one parameter combination	SMOTE,PCA,AE(for full data set of ACS) SMOTE,PCA,AE,COPULA, EFPA_HIST(for the sample of ACS data)

Evaluation of Synthetic instances:

1.Methods and Method Category used:

In this challenge, we used two different synthetic generation methodologies. In the first methodology, apply differentially private synthetic generation methods, such as the **DP Histogram** method or **DP Copula** method to the ***entire original*** dataset. The resulting generated dataset from those two methods can be considered completely differentially private and fully synthetic. In the second methodology, apply methods, such as **Autoencoders (AE), Principal Component Analysis (PCA) and Synthetic Minority Oversampling Technique (SMOTE)** on the numerical features of the original data and apply the DP mechanism (Laplace) to only the categorical features of the original data. In the second methodology, the generated dataset is, hence, deemed to be partially differentially private but fully synthetic. The main difference between the two mentioned methodologies is that *completely differentially private* techniques have the goal of allowing analysts to learn about *trends* in sensitive data, and mathematically bound the risk of revealing information specific to *individuals* within a parameter epsilon (ε), making the generated data ε-differentially private. Comparatively, with *incompletely differentially-private techniques*, we can only guarantee that categorical features of individuals have a bounded risk of revealing sensitive information, and cannot mathematically bound the privacy risk of the numerical features.

UNCLASSIFIED

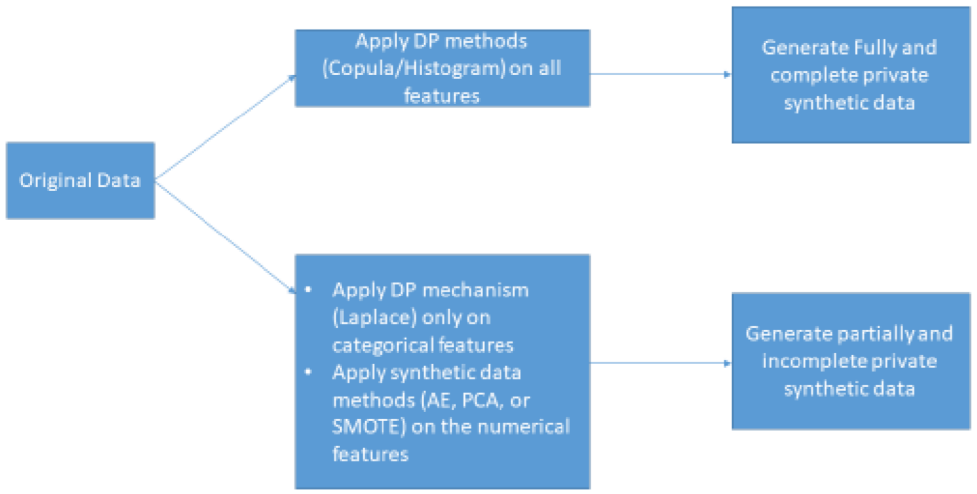

Figure 1: Methodologies and Methods used in synthetic data generation

In the case of *satgpa* data set, the data set size is small with only one categorical variable and five numerical variables. All the above mentioned five methods(DP Hist, DP Copula, PCA, SMOTE, AE) used and all the variables are synthesized. Second methodology is applied to ACS data set, as ACS data set will

require longer computation time for synthesis and evaluation because of its large size and many features. Selected few features out of all the 33 features and partial synthesized the this data set.

2. Tools used to generate the synthetic data:

R programming tool is used to create the pipeline to generate the synthetic data and evaluate the metrics as well.

3. Evaluation of the synthetic data:

Regardless of which synthetic data generation method is used, without actually comparing a generated dataset to the original it was generated from, it is hard in practise to evaluate how much utility the generated data will give, and also how much of a practical privacy risk exists. Therefore, we evaluate the practical privacy risks and the utilities of each synthetic data generation method using some measures, scores, and metrics.

3.1.Utility Metrics:

I. ***Overall summary metrics:***

- MEAN.DIFF: For each of the continuous variables , compute the mean of the original data ($\bar{x}_{i,orig}$) and that of the synthetic data ($\bar{x}_{i,synth}$) and calculate the difference $\Delta\bar{x}_i = \bar{x}_{i,orig} - \bar{x}_{i,synth}$.
- MEDIAN.DIFF:Similarly for each of the continuous variables , compute the median of the original data and that of the synthetic data, and calculate the difference.
- Similarly ,for each continuous variables, calculate the minimum difference and maximum difference between original and synthetic data.
- If these differences are too small indicates the synthetic data distributions are close to the original data.

II. ***Utility metrics(Using PMSE):***

This is a general utility approach, called Propensity Score Measure (pMSE), to measure overall distribution similarity between the synthetic and original data. The approach presented here uses the concept of propensity scores (predicted probabilities of group membership) to discriminate between the original and synthetic data. Steps involved here to calculate this metrics is

- Merge the original and synthetic data sets, adding a variable Y equal to one for all rows from the synthetic data set and equal to zero for all rows from the original data set.
- Second, for each record in the original and synthetic data, compute the probability of being in the synthetic dataset, i.e. the propensity score.
- Compare the distributions of the propensity scores in both datasets.

III. PEARSON PAIRWISE CORRELATION COEFFICIENTS:

- Calculate the pairwise Pearson correlation coefficients within the original dataset, p_"orig" , and the synthetic dataset, p_"synth" .

- Plot the difference Δp=p_"orig" -p_"synth" for each pairwise combination of continuous variables in our dataset .
- Δp will be close to zero for variables that have similar correlation coefficients in the synthetic data as they do in the original data.
- To assess the entire synthetic dataset, sum |Δp| across all pairwise combinations to get a single metric μ_"abs" that is an overall indicator of how much the pairwise relationships between variables differ.
- Smaller values of μ_"abs" indicate that the correlations between variables are, on average, representative of those in the original dataset.

IV. ***Random cuts:***
Filter by every feature and compare counts to original data

3.2. Privacy Metrics:

I. ***Full Intersections:***

The first and most basic test to measure privacy is to count the number of full intersections between the synthetic and original dataset. This is straightforward, and one simple approach to minimizing these intersections is to add some noise or to remove them entirely from the synthetic dataset in post processing. From an adversarial point of view, this doesn't give the ability to infer information about all individuals in the dataset, but could give the ability to make an accurate guess about a few individuals.

II. ***Pairwise Intersections:***

A pairwise intersection occurs if any two rounded columns in the original dataset are the same as any two rounded rows in the synthetic data. For this metric, we count all pairwise intersections in the numeric variables, within each category group.

I. ***Fuzzy matching distance:***

Pairwise intersections and duplicates only count perfect matches. However, in order to capture a "very close" pairwise match, we created another metric based on "closest point differences". The idea here is to assume an adversary knowns information about feature X1 in the original dataset, and looks to use that information to infer feature X2. One approach they can use is to look in the synthetic dataset, and find the closest value to X1 (X1_syn). Then they can use X2_syn at the same index as an approximate X2 real value. What we want to measure is the absolute difference between this X2_syn and real X2 for all features.

- Mean of Fuzzy distance:Merge the original and synthetic data sets, adding a variable Y equal to one for all rows from the synthetic data set and equal to zero for all rows from the original data set.
- Mean of Inverse Fuzzy distance: invert the distance sum for each observation, and take a mean of the inverses: 1/(pmax(abs(X2_syn-X2),k)). Note that the pmax in the denominator is required because for points that are perfectly inferred, this distance sum will be exactly 0, making the inverse infinity. For the results in this report, we take the k=1, meaning that the results will effectively compute the percentage of points that are

within 1 unit of the original. By setting the parameter k higher, we can compute the number of points “within k units”, and this is useful in case for a particular application “close” is defined differently.

In summary, a small mean of inverse fuzzy distance indicates that there are not many close matches between the synthetic and original datasets, while large mean inverse fuzzy distance indicate there are many close matches between the synthetic and original dataset.

4.Results:

Here are the results of utility measures evaluation for the satgpa data set.

The difference in the means between original and synthetic features is shown here with PCA and SMOTE giving results closest to zero, indicating that these two methods best preserve aggregate statistics. AE consistently skews toward higher means. COP and EFPA_HIST vary by feature. Note that NULL is non zero as this is a random sample of the full original data and therefore may differ in mean. As epsilon increases, the y axis scale increases, illustrating how the privacy trades off with accuracy in reproducing properties of the dataset.

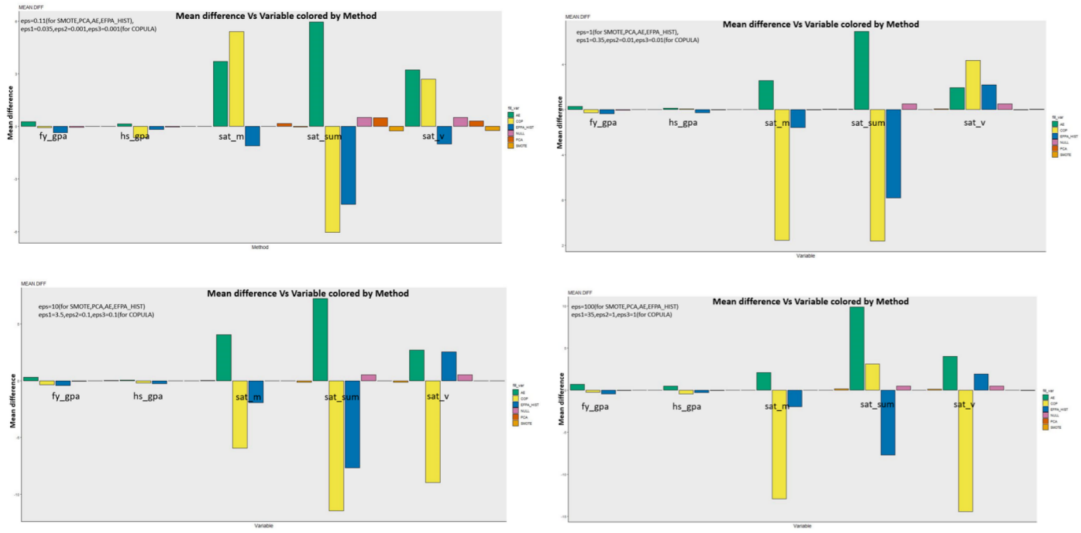

Figure 2:Difference of Mean of Original and Mean of Synthetic data for continuous variables Vs Methods

For the difference in medians, COP and EFPA_HIST are the farthest from the median of the original data. The y axis scale is more stable compared to the mean case, so that the median appears less subject to fluctuations relative to the privacy budget. PCA and SMOTE perform well.

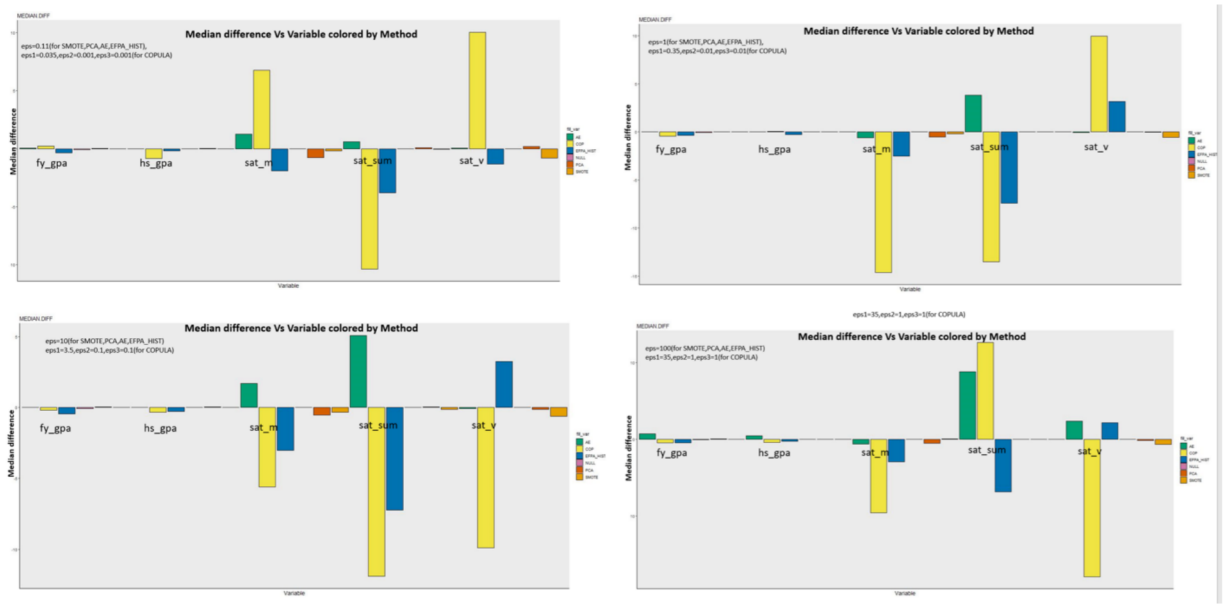

Figure 3:Difference of Median of Original and Median of Synthetic data for continuous variables Vs Methods

When epsilon is low (0.1) all the methods fare equally badly at reproducing the maximum value of the original. COP is among the worst in terms of reproducing the maximum when epsilon is greater than 1. PCA, SMOTE and AE are better.

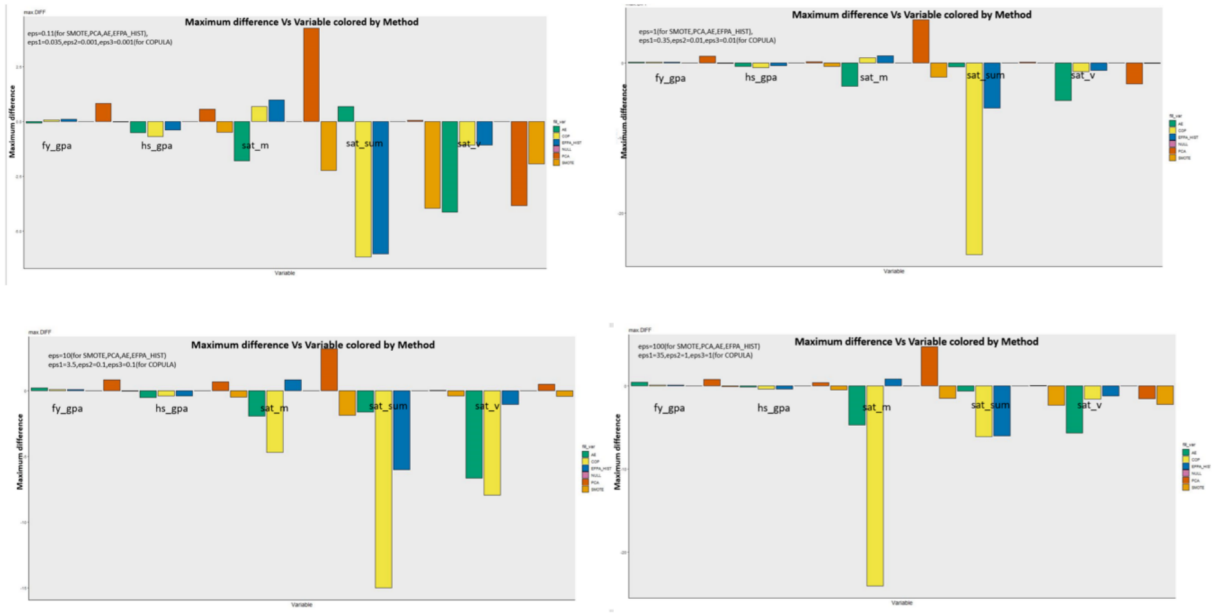

Figure 4:Difference of Maximum of Original and Maximum of Synthetic data for continuous variables Vs Methods

For the minimum of each feature, AE performs the worst at all epsilon values and for all features.

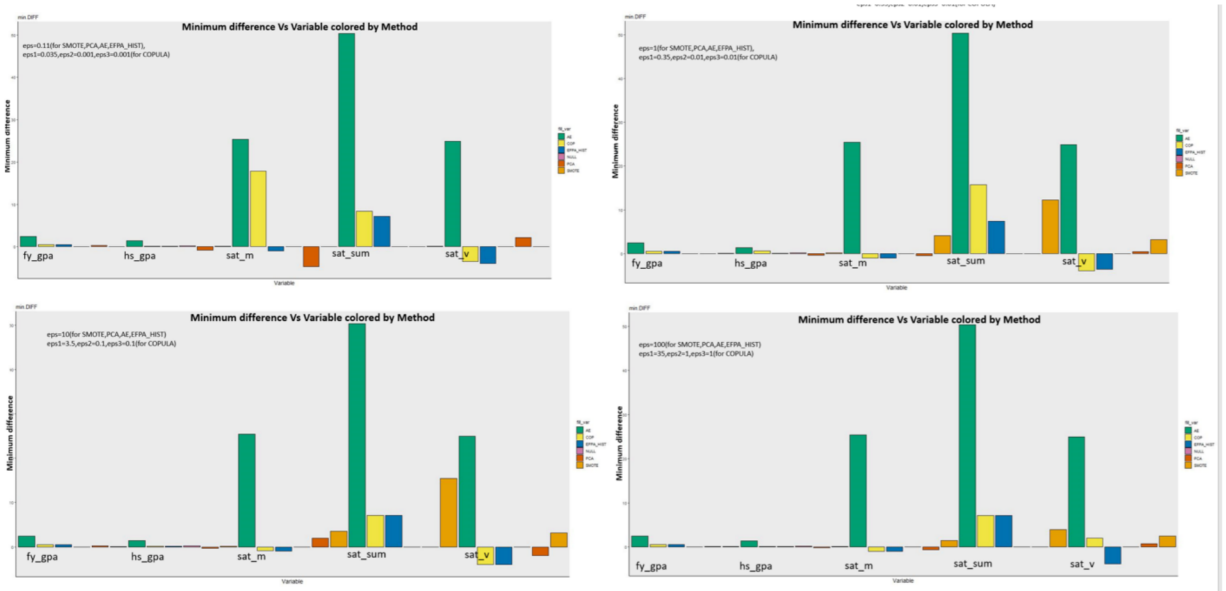

Figure 5: Difference of Minimum of Original and Minimum of Synthetic data for continuous variables Vs Methods

The last aggregate statistic we show is difference in the fifth percentiles. PCA is closest to the original, then SMOTE. COP and EFPA_HIST are similar but perform more poorly on this and AE is the farthest from the original.

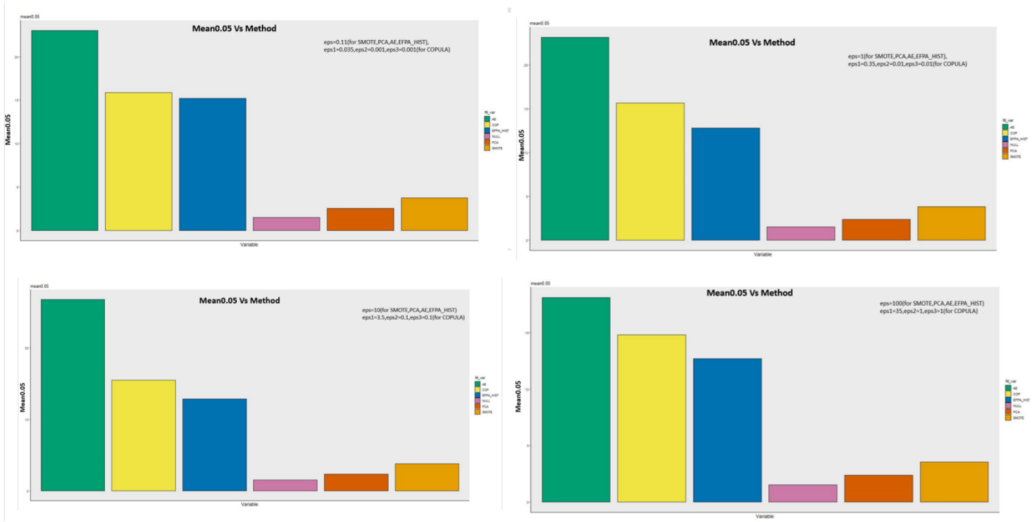

Figure 6: Mean0.05 Vs Methods

The utility based on logistic regression classifier shows that the synthetic data generated using PCA and SMOTE is difficult to tell apart from the original as the score is close to zero. AE, COP and EFPA_HIST all score 0.25 meaning they can be readily differentiated from the original.

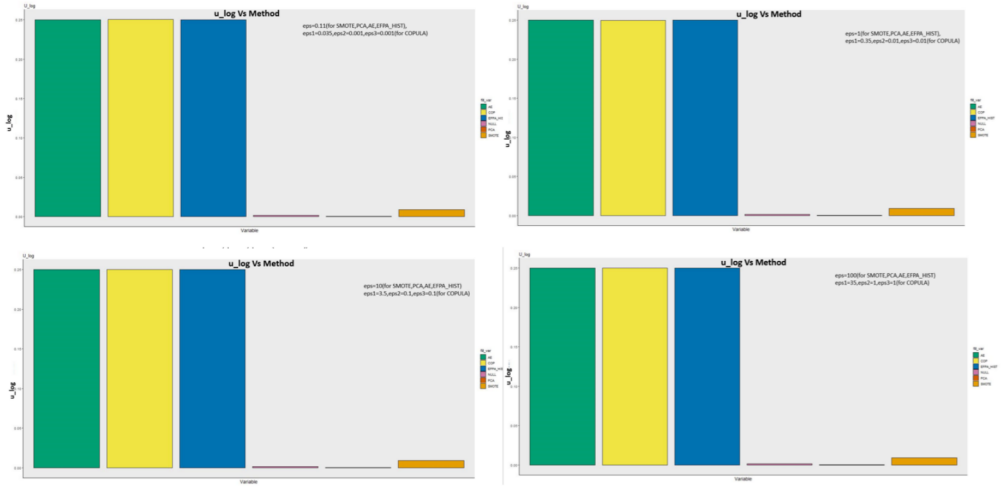

Figure 7: U_log Vs Methods

The utility based on decision tree classifier is more nuanced: SMOTE is the closest to the original data by this metric followed by EFPA_HIST. The remaining metrics score between 0.2 and 0.25, indicating that the classifier can almost perfectly distinguish them from the original.

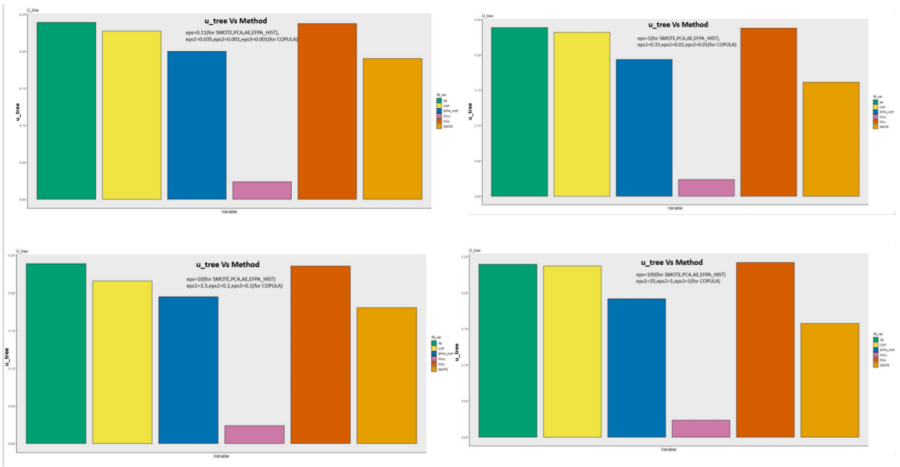

Figure 8: u_tree Vs Methods

To capture whether pairwise correlations are well preserved between numerical features, we compute the absolute sum of pairwise correlation differences with respect to the original data. A value close to

zero indicates that the correlations are close to those in the original data. PCA and SMOTE preserve these properties best.

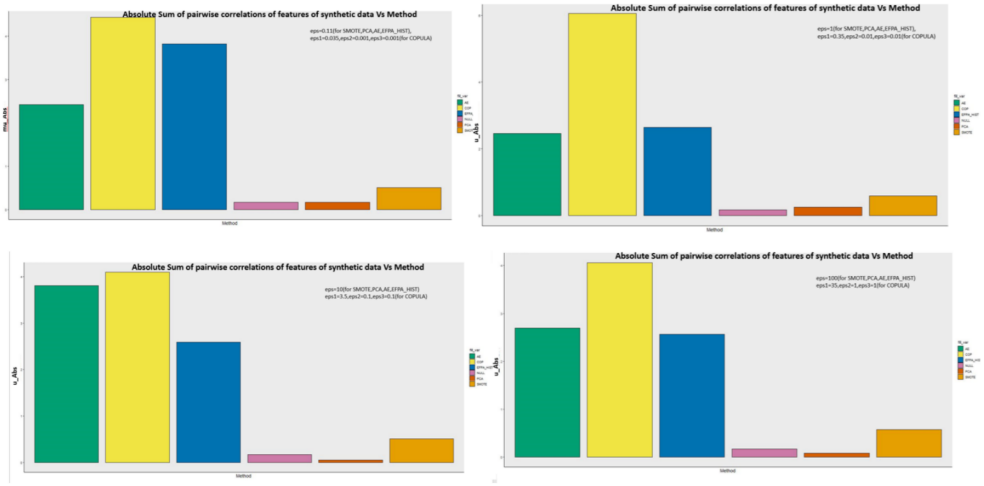

Figure 9: ***Absolute Sum of pairwise correlations of features of synthetic data difference with Original Vs Method***

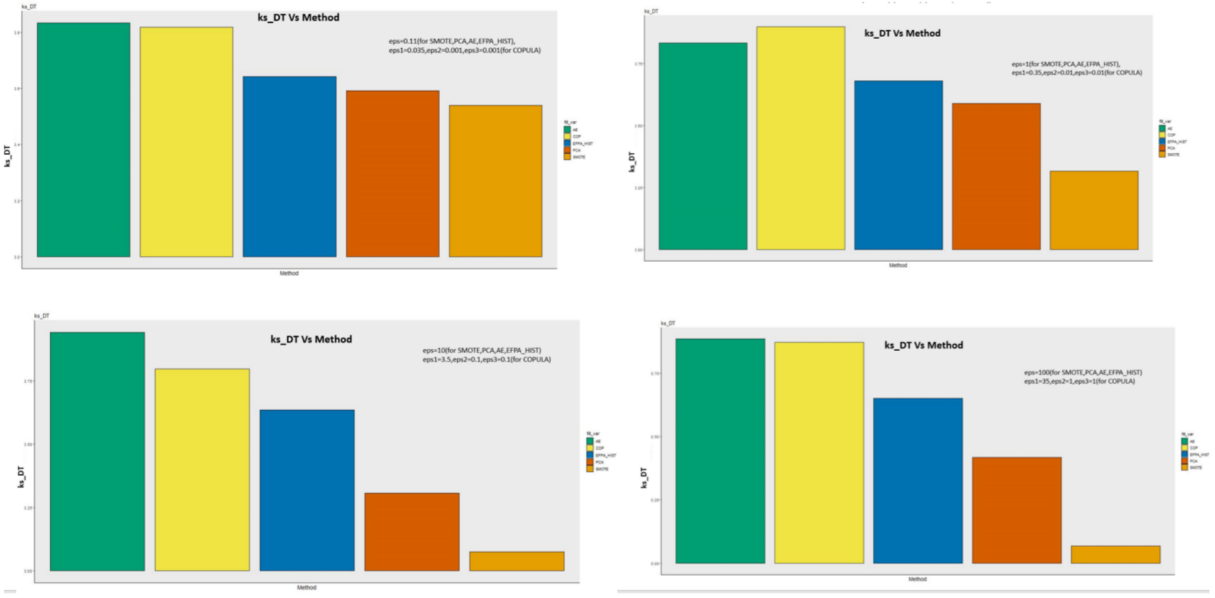

Figure 10: ks_DT Vs Methods

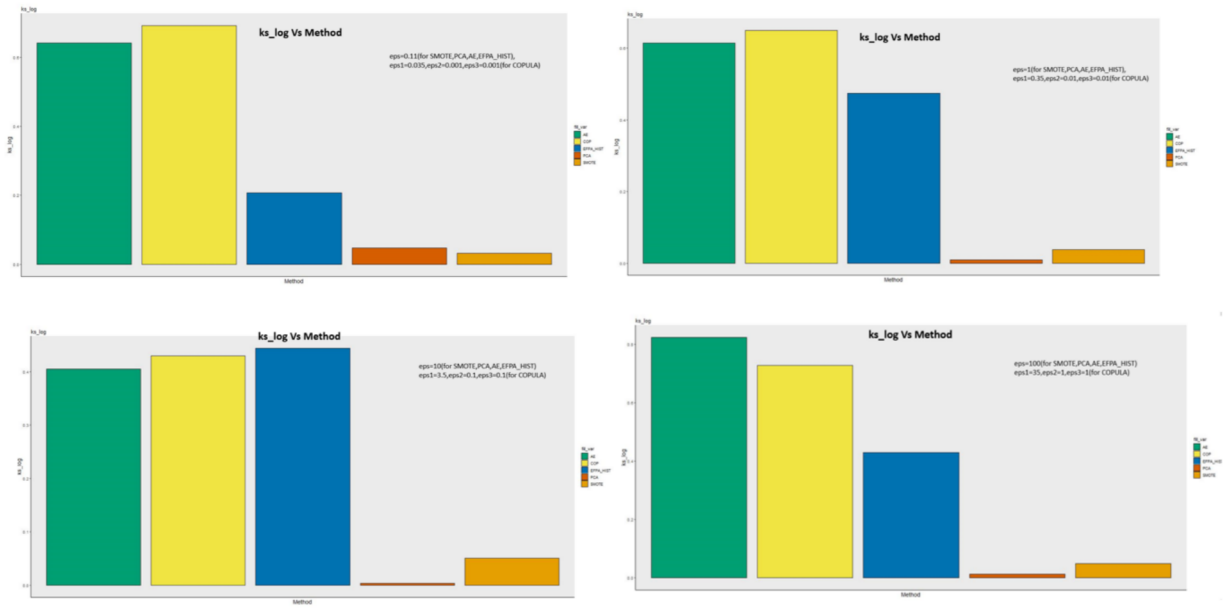

Figure 11:ks_log Vs Methods

Privacy metrics results for satgpa data set:

The full intersections show the number of cases in which rows from the original data are reproduced in the synthetic data. The AE as well as the two fully differentially private methods COP and EFPA_HIST have no exact replicas of original individuals.

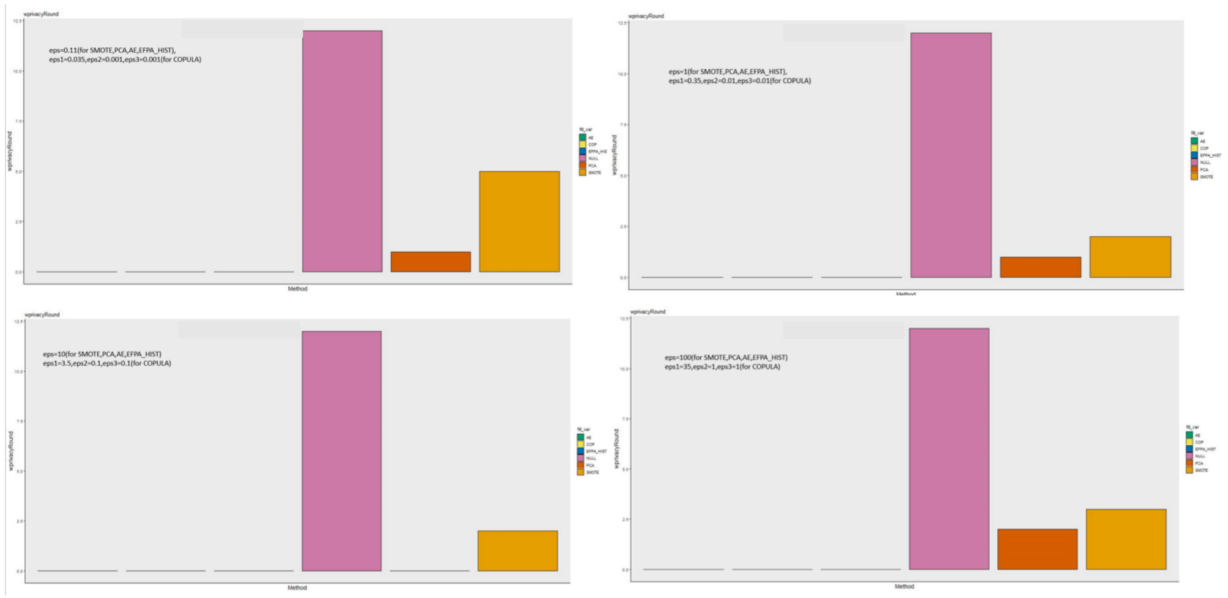

Figure 12: Full Intersections Vs Methods

Pairwise intersections are cases where two numerical values in a row are reproduced from the original data when rounded to unit. AE, COP and EFPA_HIST are more private.

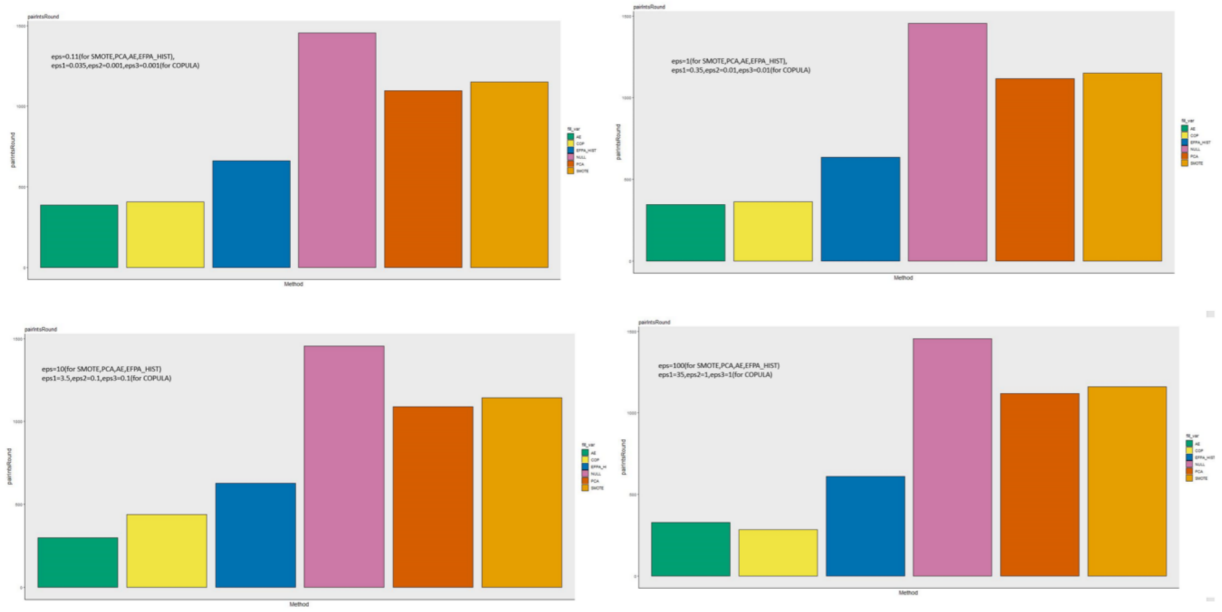

Figure 13:Pairwise Intersections Vs Method

Duplicates is another way of counting the intersections and the results are consistent with the previous figure.

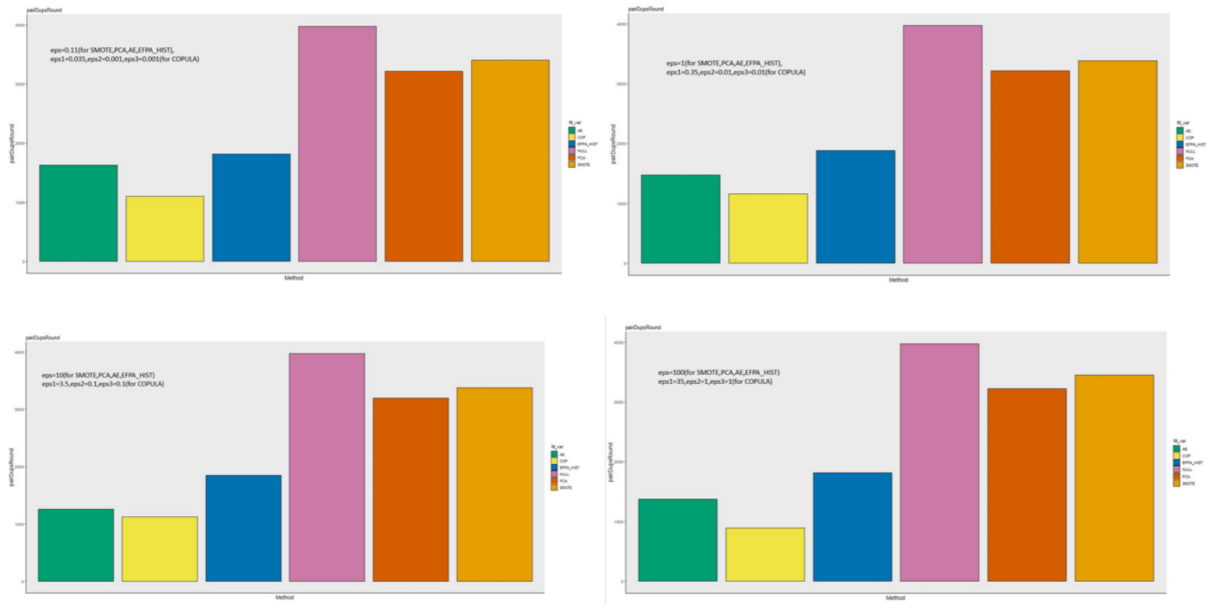

Figure 14: Duplicates Vs Methods

The inverse fuzzy matching distance is a measure of how easily one numerical feature can be inferred from another. Values close to 1 mean the results are on par with the original data. Using this metric, all the methods perform similarly but COP and EFPA_HIST are slightly more secure toward this type of inference.

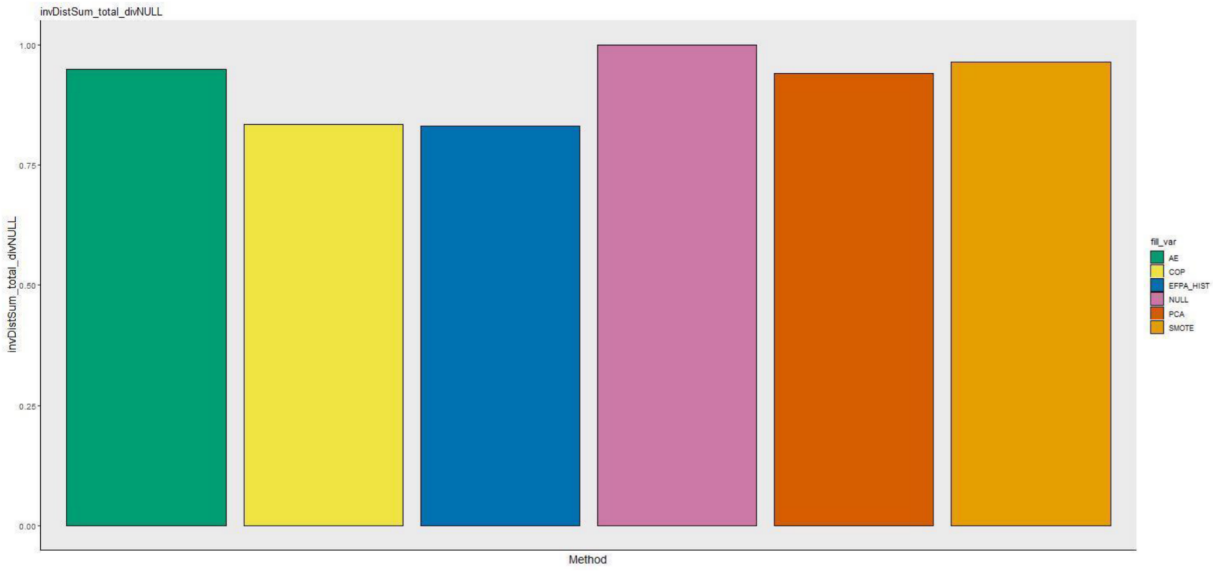

Figure 15:Inverse Fuzzy matching distance Vs methods

Results for ACS data set:

For the larger ACS data, we ran only the partially differentially private methods. The pairwise correlations were closest to the original when using AE. However, these are all on a scale below 0.03, a very small difference overall.

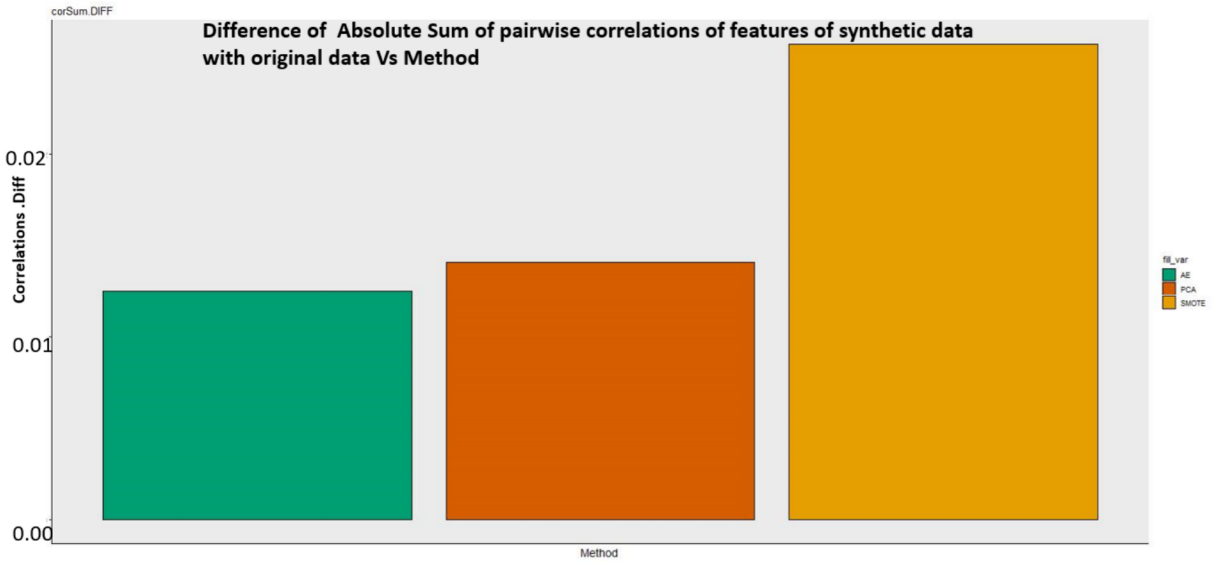

Figure 16: Difference of Absolute Sum of pairwise correlations of features of synthetic data with original data Vs Method

This shows the sum of absolute differences with respect to the original across feature means to summarize aggregate results. The methods are all comparable when it comes to preserving the feature means.

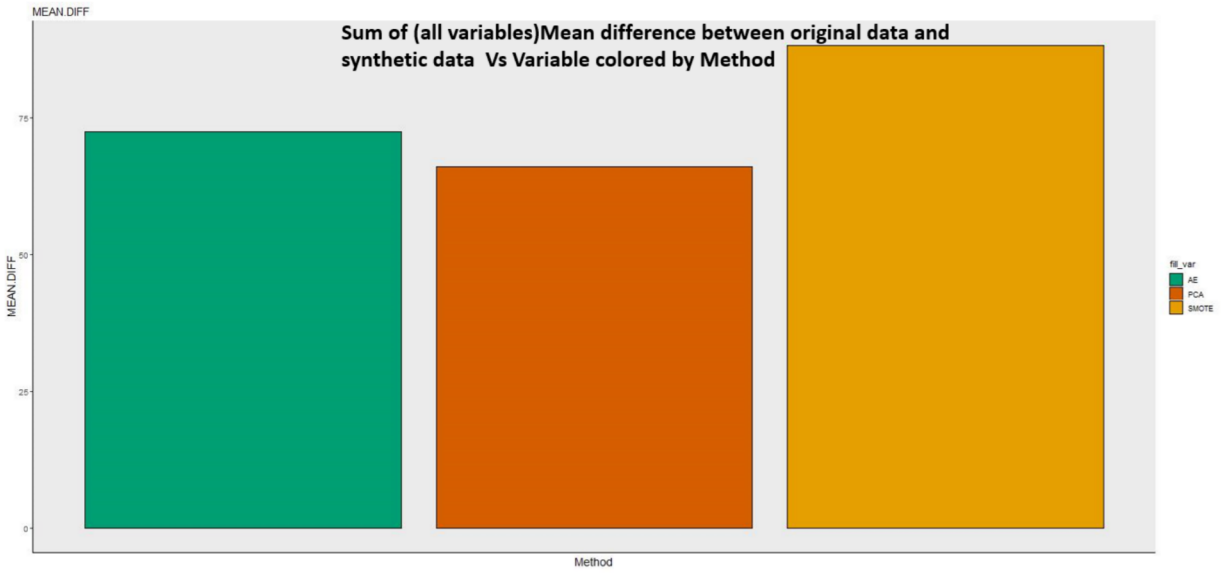

Figure 17:Sum of (all variables)Mean difference between original data and synthetic data Vs Variable colored by Method

For this and the next figure, we can see that the utility should be high for all three methods as they produce data nearly indistinguishable from the original according to the scores many orders of magnitude below 0.25.

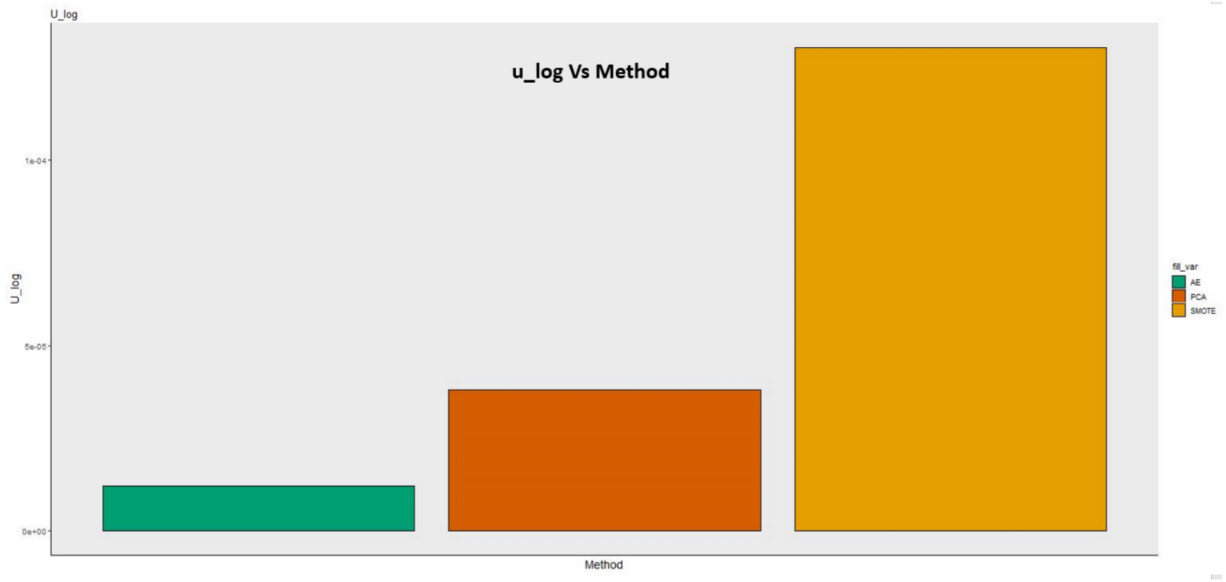

Figure 18:u_log Vs Method

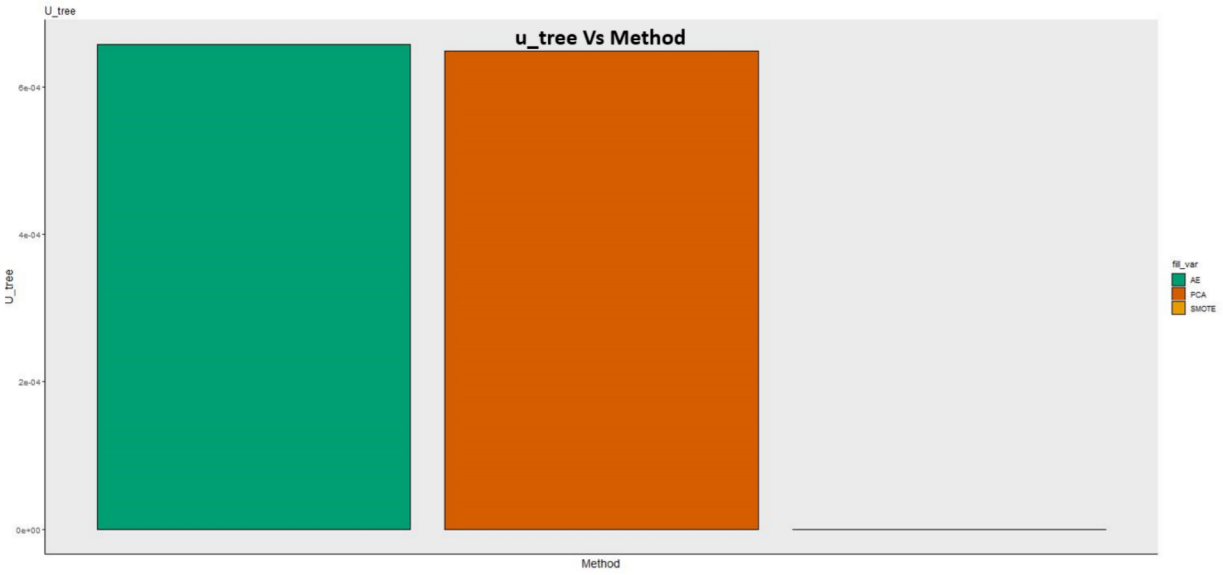

Figure 19:u_tree Vs Method