NIST CRC Meta Report
SDV

Report created on: May 19, 2023 23:46:53

Created with SDNIST v2.2.1

Motivation

The Synthetic Data Vault (SDV) is a user-friendly, open source python library of non-differentially private synthetic data approaches.

In our archive we have samples from several SDV methods.

Gaussian Copula is a purely statistical modeling approach. It first fits Gaussian Copula functions (think multi-dimensional, skewed bell curves) to the target distribution, this gives it an approximation of the shape of the target distribution. It uses these to sample new data that fits that shape.

Other methods from SDV are neural network based. Because they aren't differentially private (and thus don't need to limit how they interact with the target data), the number of iterations they use for training (epochs) can have a significant impact on their output quality. FastML and CTGan are both based on Generalized Adversarial Networks (GAN) -- the first is simple and runs quick, the latter runs more slowly and is designed to work better on tabular data. Copula GAN is another GAN technique (an new, experimental one). By contrast, TVAE is a different type of neural network approach-- Variational Auto Encoder. It transforms records into a new feature space (encoding) that improves modeling of feature relationships, and then uses this space to generate new records.

You can find these and other synthesis methods in the Synthetic Data Vault library here: [sdv]

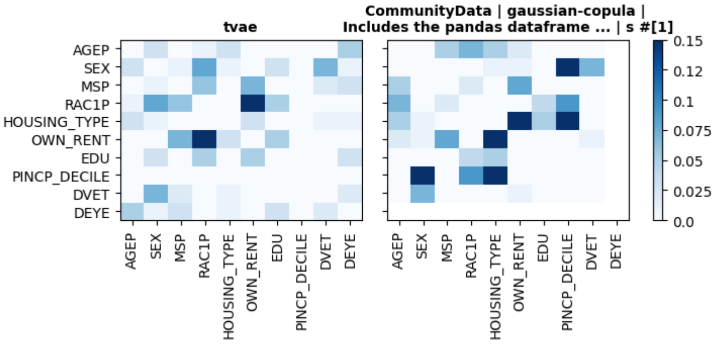
Comparisons

Correlation Comparison:

The Pearson Correlation difference was a popular utility metric during the HLG-MOS Synthetic Data Test Drive. Note that darker highlighting indicates pairs of features whose correlations were not well preserved by the deidentified data.

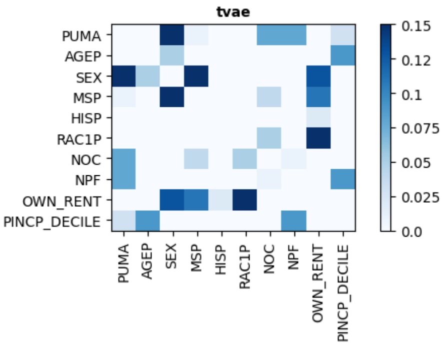
Feature Set: demographic-focused | Target Dataset: ma2019:

Features: ['DEYE', 'SEX', 'HOUSING_TYPE', 'OWN_RENT', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'AGEP']
Feature Space (possible combinations): 227,026,800



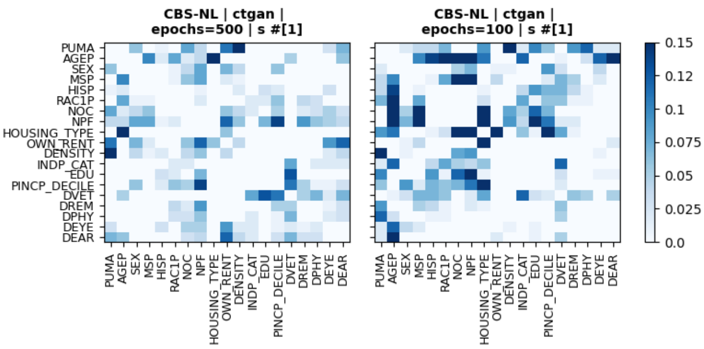
Feature Set: family-focused | Target Dataset: ma2019:

Features: ['SEX', 'OWN_RENT', 'HISP', 'MSP', 'RAC1P', 'PINCP_DECILE', 'PUMA', 'AGEP', 'NOC', 'NPF', 'POVPIP']
Feature Space (possible combinations): 64,449,000



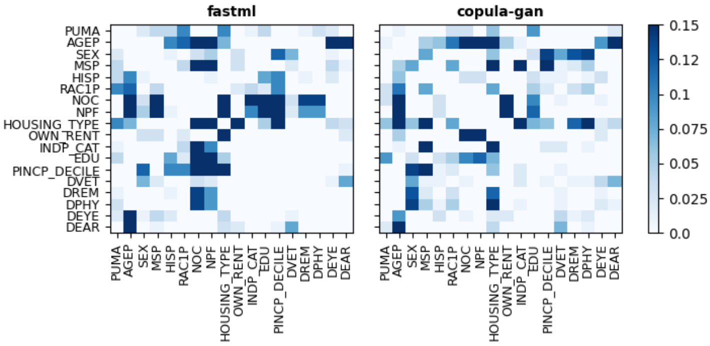
Feature Set: all-features | Target Dataset: ma2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'DENSITY', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000



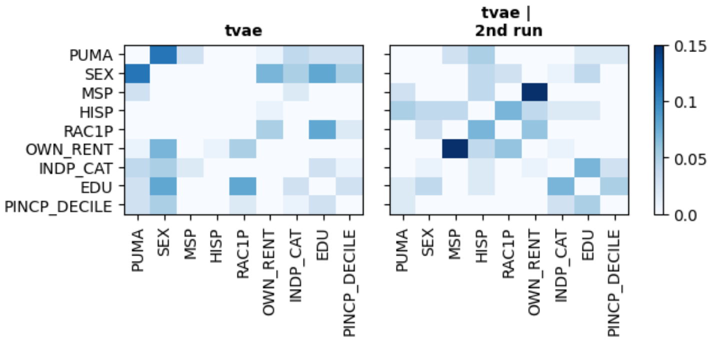
Feature Set: custom-features-23 | Target Dataset: ma2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000



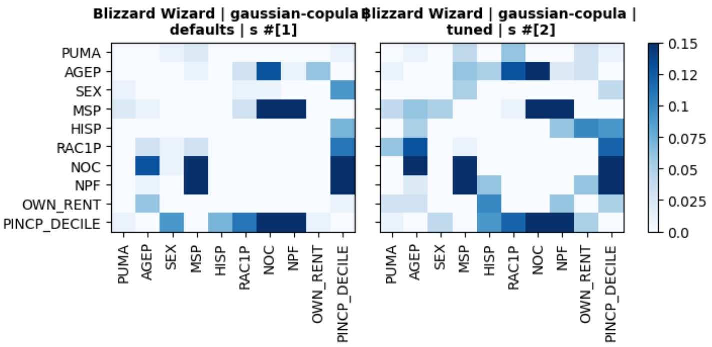
Feature Set: industry-focused | Target Dataset: national2019:

Features: ['SEX', 'OWN_RENT', 'HISP', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA']
Feature Space (possible combinations): 167,567,400



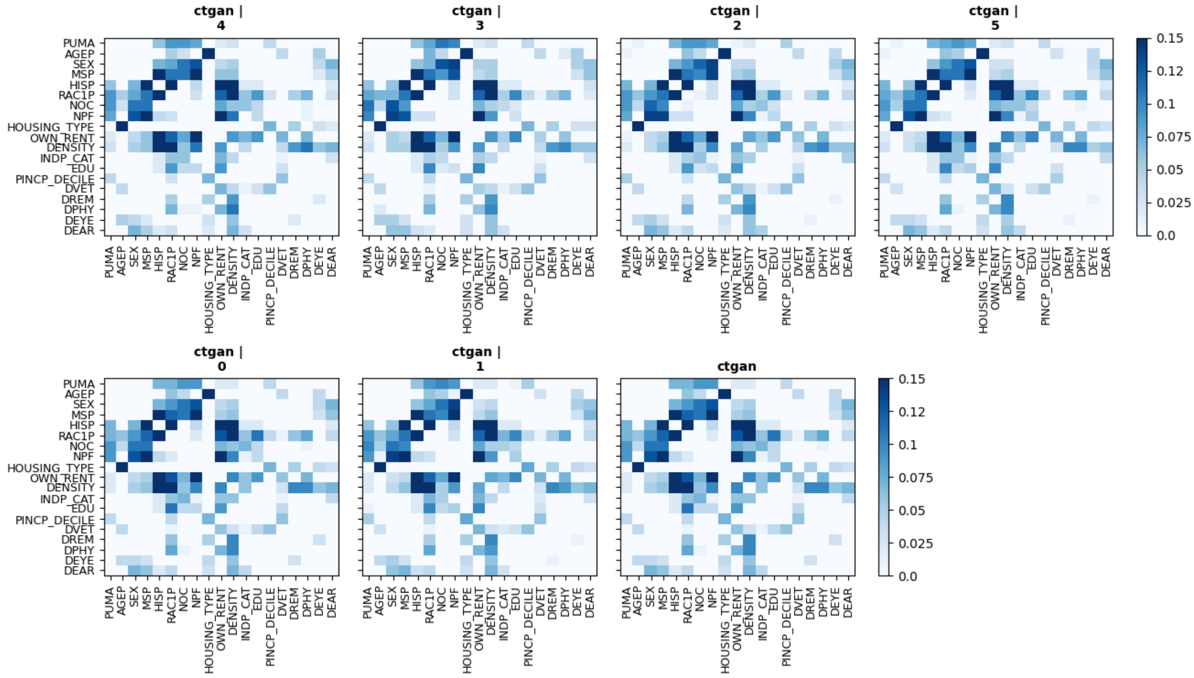
Feature Set: family-focused | Target Dataset: national2019:

Features: ['SEX', 'OWN_RENT', 'HISP', 'MSP', 'RAC1P', 'PINCP_DECILE', 'PUMA', 'AGEP', 'NOC', 'NPF', 'POVPIP']
Feature Space (possible combinations): 64,449,000



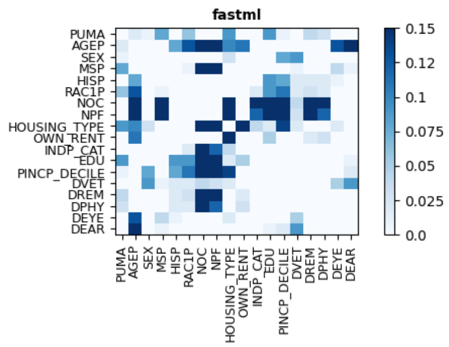
Feature Set: all-features | Target Dataset: national2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'DENSITY', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000



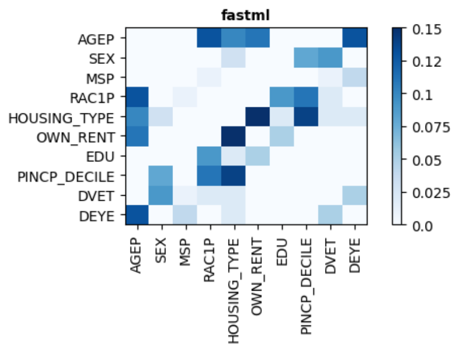
Feature Set: custom-features-23 | Target Dataset: national2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000



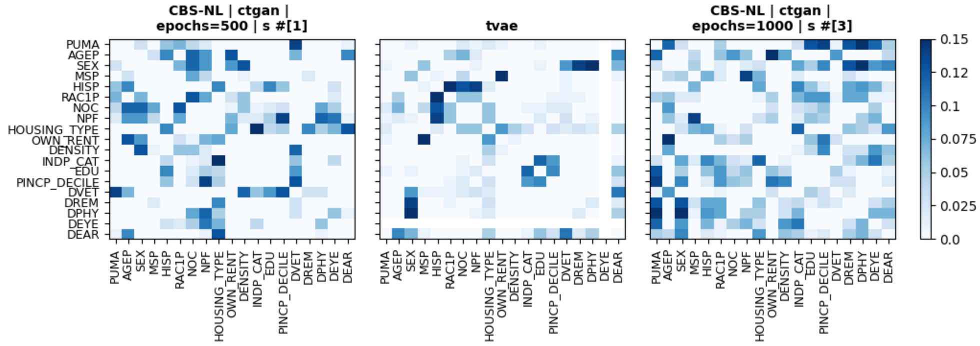
Feature Set: demographic_focused | Target Dataset: national2019:

Features: ['DEYE', 'SEX', 'HOUSING_TYPE', 'OWN_RENT', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'AGEP']
Feature Space (possible combinations): 227,026,800



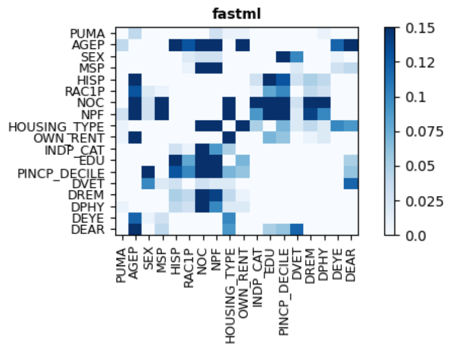
Feature Set: all-features | Target Dataset: tx2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'DENSITY', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000



Feature Set: custom-features-23 | Target Dataset: tx2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000



Unique Exact Matches Comparison:

This is a count of unique records in the target data that were exactly reproduced in the deidentified data. Because these records were unique outliers in the target data, and they still appear unchanged in the deidentified data, they are potentially vulnerable to reidentification.

Feature Set: demographic-focused | Target Dataset: ma2019:

Features: ['DEYE', 'SEX', 'HOUSING_TYPE', 'OWN_RENT', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'AGEP']
Feature Space (possible combinations): 227,026,800

Number of Unique Records in Target Data: 4268 (55.91%)

Variant	Records Matched In Target Data	Percent Records Matched In Target Data
tvae	556	7.28
CommunityData	gaussian-copula	Includes the pandas dataframe ...	s # [1]	22	0.29

Feature Set: family-focused | Target Dataset: ma2019:

Features: ['SEX', 'OWN_RENT', 'HISP', 'MSP', 'RAC1P', 'PINCP_DECILE', 'PUMA', 'AGEP', 'NOC', 'NPF', 'POVPIP']
Feature Space (possible combinations): 64,449,000

Number of Unique Records in Target Data: 6435 (84.29%)

Variant	Records Matched In Target Data	Percent Records Matched In Target Data
tvae	387	5.07

Feature Set: all-features | Target Dataset: ma2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'DENSITY', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000

Number of Unique Records in Target Data: 7626 (99.9%)

Variant	Records Matched In Target Data	Percent Records Matched In Target Data
CBS-NL	ctgan	epochs=500	s #[1]	0	0.0
CBS-NL	ctgan	epochs=100	s #[1]	0	0.0

Feature Set: custom-features-23 | Target Dataset: ma2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000

Number of Unique Records in Target Data: 7626 (99.9%)

Variant	Records Matched In Target Data	Percent Records Matched In Target Data
fastml	0	0.0
copula-gan	0	0.0

Feature Set: industry-focused | Target Dataset: national2019:

Features: ['SEX', 'OWN_RENT', 'HISP', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA']
Feature Space (possible combinations): 167,567,400

Number of Unique Records in Target Data: 16132 (59.19%)

Variant	Records Matched In Target Data	Percent Records Matched In Target Data
tvae	1412	5.18
tvae	2nd run	1166	4.28

Feature Set: family-focused | Target Dataset: national2019:

Features: ['SEX', 'OWN_RENT', 'HISP', 'MSP', 'RAC1P', 'PINCP_DECILE', 'PUMA', 'AGEP', 'NOC', 'NPF', 'POVPIP']
Feature Space (possible combinations): 64,449,000

Number of Unique Records in Target Data: 23908 (87.73%)

Variant	Records Matched In Target Data	Percent Records Matched In Target Data
Blizzard Wizard | gaussian-copula | defaults | s #[1]	0	0.0
Blizzard Wizard | gaussian-copula | tuned | s #[2]	1	0.0

Feature Set: all-features | Target Dataset: national2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'DENSITY', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000

Number of Unique Records in Target Data: 27159 (99.66%)

Variant	Records Matched In Target Data	Percent Records Matched In Target Data
ctgan | 4	0	0.0
ctgan | 3	0	0.0
ctgan | 2	0	0.0
ctgan | 5	0	0.0
ctgan | 0	0	0.0
ctgan | 1	0	0.0
ctgan	0	0.0

Feature Set: custom-features-23 | Target Dataset: national2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000

Number of Unique Records in Target Data: 27159 (99.66%)

Variant	Records Matched In Target Data	Percent Records Matched In Target Data
fastml	0	0.0

Feature Set: demographic_focused | Target Dataset: national2019:

Features: ['DEYE', 'SEX', 'HOUSING_TYPE', 'OWN_RENT', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'AGEP']
Feature Space (possible combinations): 227,026,800

Number of Unique Records in Target Data: 14918 (54.74%)

Variant	Records Matched In Target Data	Percent Records Matched In Target Data
fastml	687	2.52

Feature Set: all-features | Target Dataset: tx2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'DENSITY', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000

Number of Unique Records in Target Data: 9260 (99.83%)

Variant	Records Matched In Target Data	Percent Records Matched In Target Data
CBS-NL | ctgan | epochs=500 | s #[1]	0	0.0
tvae	0	0.0
CBS-NL | ctgan | epochs=1000 | s #[3]	0	0.0

Feature Set: custom-features-23 | Target Dataset: tx2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000

Number of Unique Records in Target Data: 9260 (99.83%)

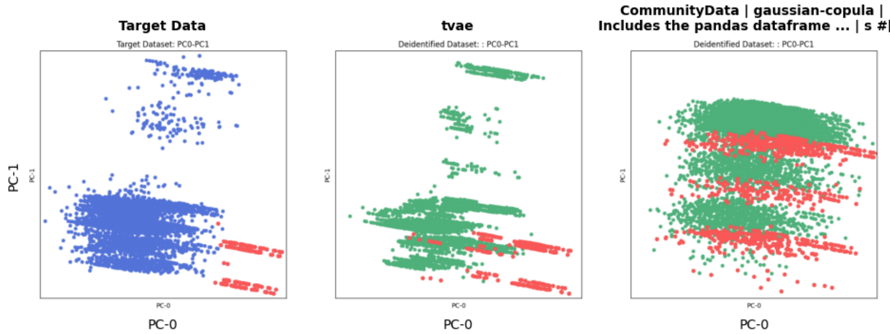
Variant	Records Matched In Target Data	Percent Records Matched In Target Data
fastml	0	0.0

PCA Comparison: (PC-0 & PC-1) with highlighted MSP-N (AGE < 15):

This is another approach for visualizing where the distribution of the deidentified data has shifted away from the target data. In this approach, we begin by using Principle Component Analysis to find a way of representing the target data in a lower dimensional space (in 5 dimensions rather than the full 22 dimensions of the original feature space). Descriptions of these new five dimensions (components) are given in the components table; the components will change depending on which target data set youâ€™re using. Five dimensions are better than 22, but we actually want to get down to two dimensions so we can plot the data on simple (x,y) axesâ€“ the plots below show the data across each possible pair combination of our five components. You can compare how the shapes change between the target data and the deidentified data, and consider what that might mean in light of the component definitions. This is a relatively new visualization metric that was introduced by the IPUMS International team during the HLG-MOS Synthetic Data Test Drive.

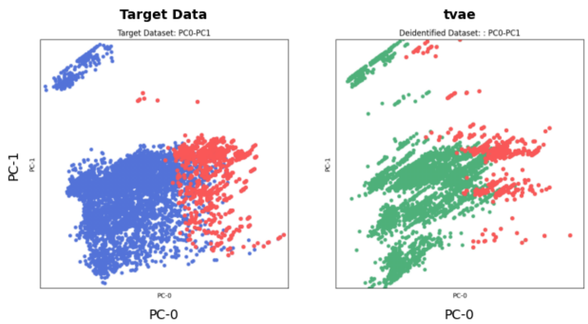
Feature Set: demographic-focused | Target Dataset: ma2019:

Features: ['DEYE', 'SEX', 'HOUSING_TYPE', 'OWN_RENT', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'AGEP']
Feature Space (possible combinations): 227,026,800



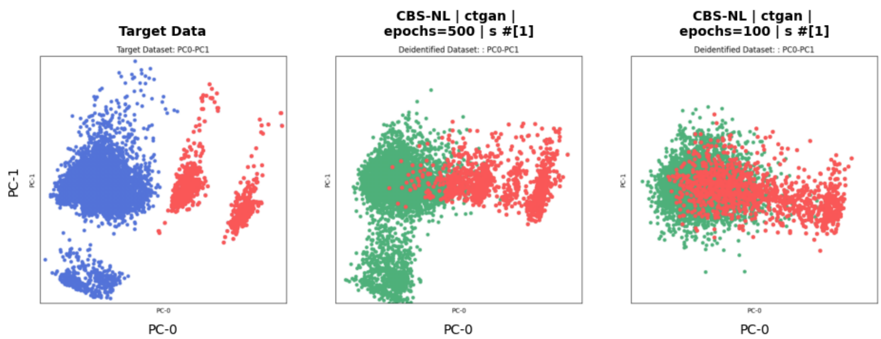
Feature Set: family-focused | Target Dataset: ma2019:

Features: ['SEX', 'OWN_RENT', 'HISP', 'MSP', 'RAC1P', 'PINCP_DECILE', 'PUMA', 'AGEP', 'NOC', 'NPF', 'POVPIP']
Feature Space (possible combinations): 64,449,000



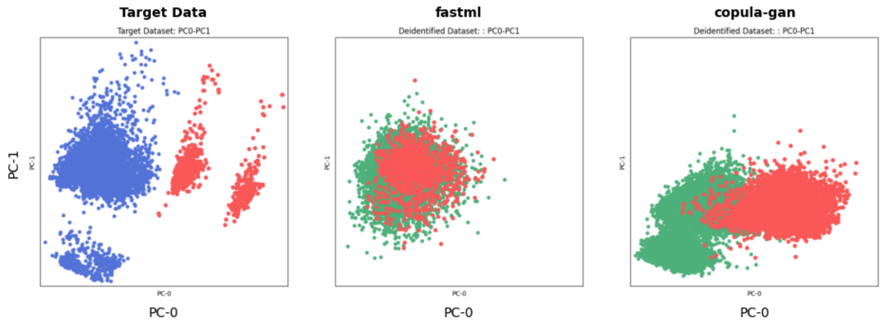
Feature Set: all-features | Target Dataset: ma2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'DENSITY', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000



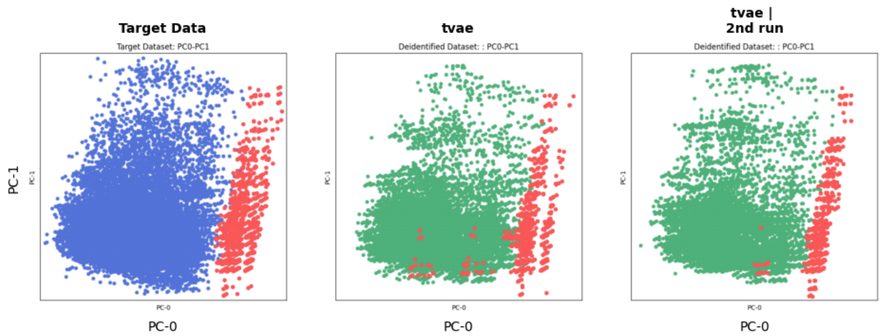
Feature Set: custom-features-23 | Target Dataset: ma2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000



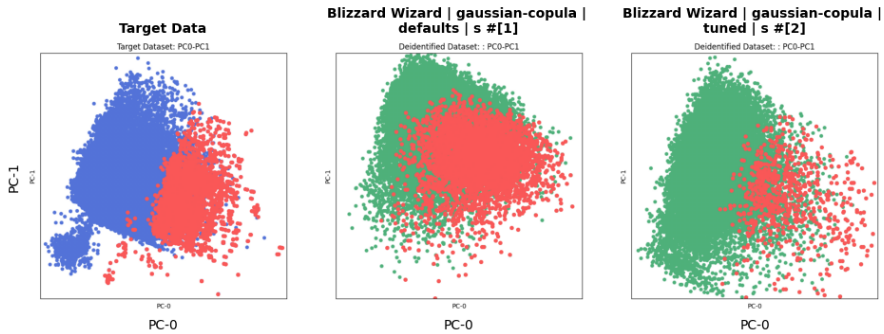
Feature Set: industry-focused | Target Dataset: national2019:

Features: ['SEX', 'OWN_RENT', 'HISP', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA']
Feature Space (possible combinations): 167,567,400



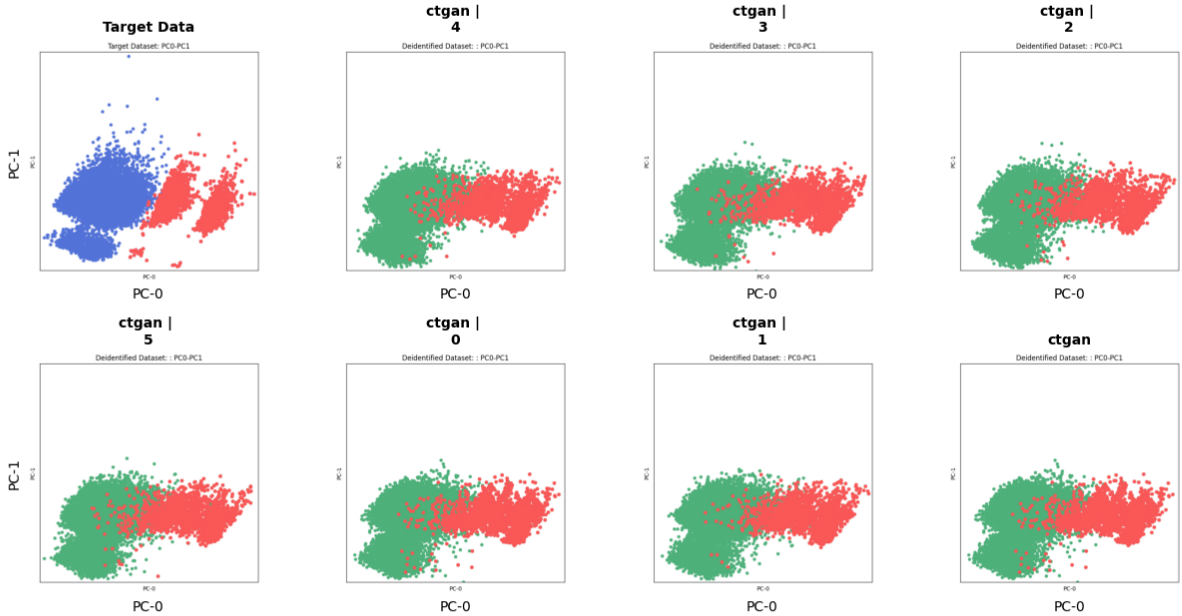
Feature Set: family-focused | Target Dataset: national2019:

Features: ['SEX', 'OWN_RENT', 'HISP', 'MSP', 'RAC1P', 'PINCP_DECILE', 'PUMA', 'AGEP', 'NOC', 'NPF', 'POVPIP']
Feature Space (possible combinations): 64,449,000



Feature Set: all-features | Target Dataset: national2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'DENSITY', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000

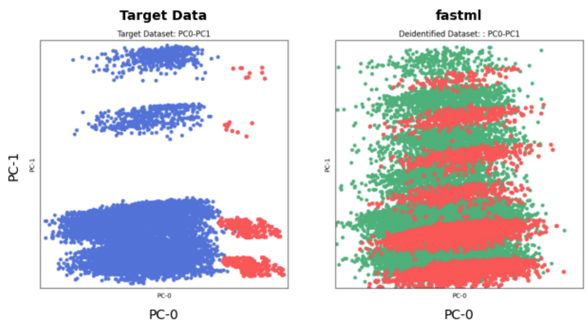


Feature Set: custom-features-23 | Target Dataset: national2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000

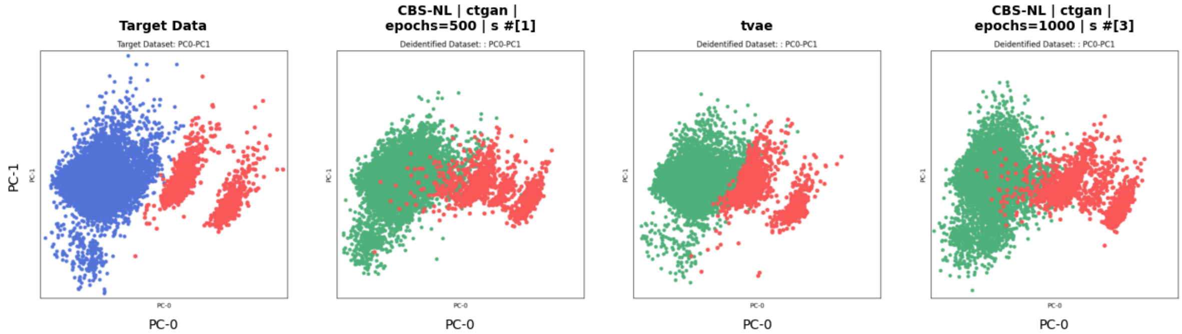
Feature Set: demographic_focused | Target Dataset: national2019:

Features: ['DEYE', 'SEX', 'HOUSING_TYPE', 'OWN_RENT', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'AGEP']
Feature Space (possible combinations): 227,026,800



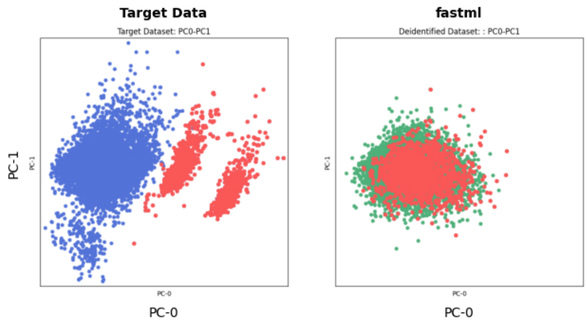
Feature Set: all-features | Target Dataset: tx2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'DENSITY', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000



Feature Set: custom-features-23 | Target Dataset: tx2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000



Regression Comparison: White Men Data:

Linear regression is a fundamental data analysis technique that condenses a multi-dimensional data distribution down to a one dimensional (line) representation. It works by finding the line that sits in the 'middle' of the data, in some sense-- it minimizes the total distance between the points of the data and the line. There are more advanced forms of regression, but here we're focusing on the simplest case-- we fit a simple straight line to the data, getting the slope and y-intercept value of that line.

For this metric we're just looking at data from adults (AGEP > 15) and we're only considering the distribution of the data across two features:

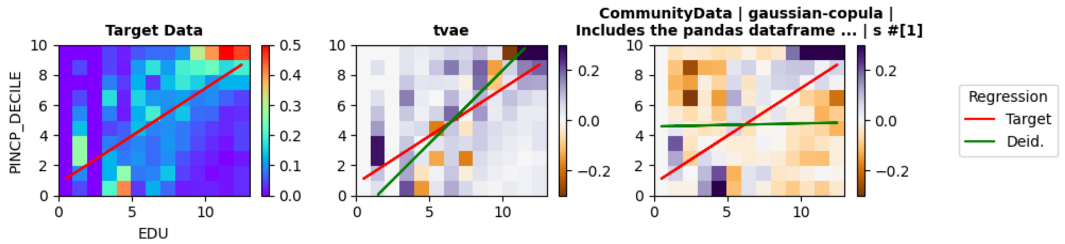
- EDU: The highest education level this individual has attained, ranging from 1 (elementary school) to 12 (PhD). See Appendix of this report for the full list of code values.
- PINCP_DECILE: The individual's income decile relative to their PUMA. This helps us account for differences in cost of living across the country. If an individual makes a moderate income but lives in a very low income area, they may have a high value for PINCP_DECILE indicating that they have a high income for their PUMA).

The basic idea is that higher values of EDU should lead to higher values of PINCP_DECILE, and this is broadly true. However, it is known that the relationship between EDU and PINCP_DECILE is different for different demographic subgroups. The heatmaps in the left column below show the density distribution of the true data for each subgroup, normalized by education category (so the density values in each column sum to 1; note that when a cell in the heatmap contains too few people (< 20), it is left blank; its not expected that the deidentified data will match the original distribution precisely). The regression line is drawn in red over the heatmap, so you can see the relationship between the target data distribution and its linear regression analysis. In the right column for each subgroup we show how the deidentified data's regression line compares to the target data's regression line, along with a heatmap of the density differences between the two distributions. Redder areas are where the deidentified data has created too many people, bluer areas are where it's created too few people.

We've broken this metric down into demographic subgroups so we can see not only how well the privacy techniques preserve the overall relationship between these features, but also whether they preserve how that overall relationship is built up from the different relationships that hold at each major demographic subgroup. It's important that deidentification techniques preserve these distinct subgroup patterns for analysis.

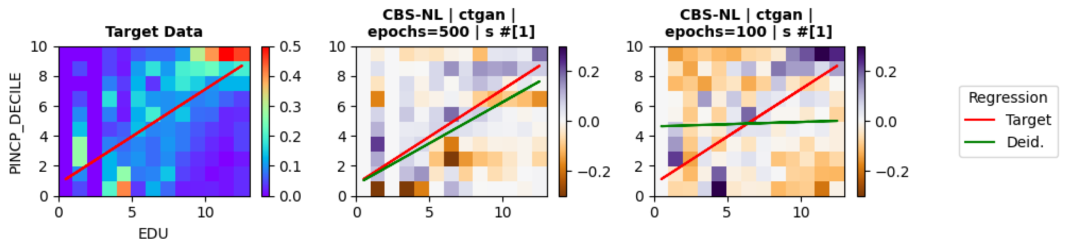
Feature Set: demographic-focused | Target Dataset: ma2019:

Features: ['DEYE', 'SEX', 'HOUSING_TYPE', 'OWN_RENT', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'AGEP']
Feature Space (possible combinations): 227,026,800



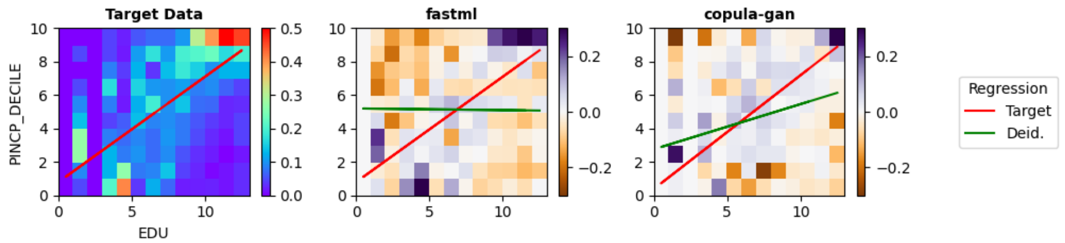
Feature Set: all-features | Target Dataset: ma2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'DENSITY', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000



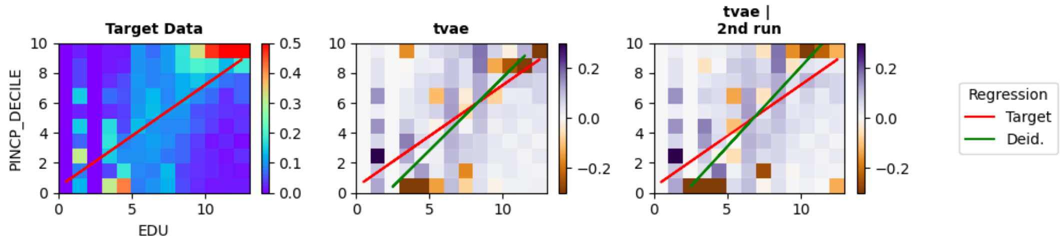
Feature Set: custom-features-23 | Target Dataset: ma2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000



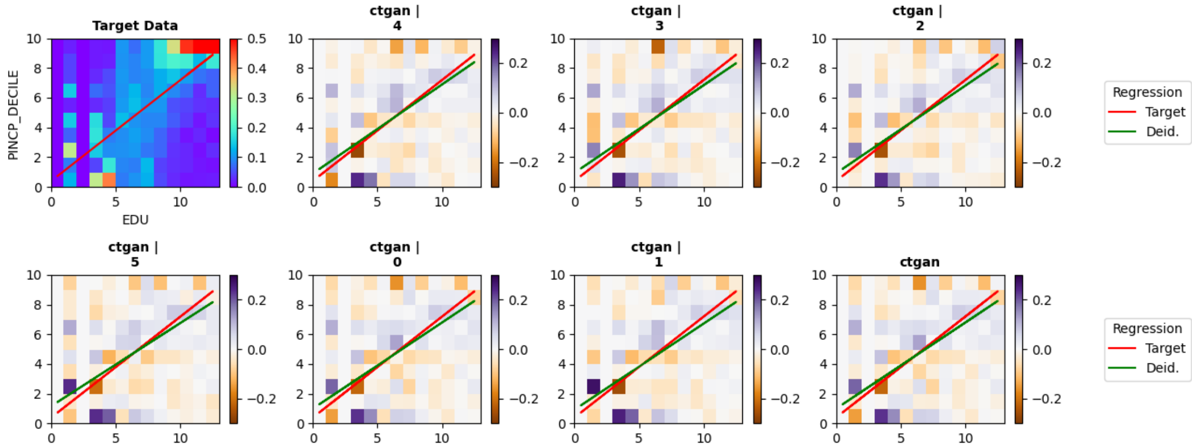
Feature Set: industry-focused | Target Dataset: national2019:

Features: ['SEX', 'OWN_RENT', 'HISP', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA']
Feature Space (possible combinations): 167,567,400



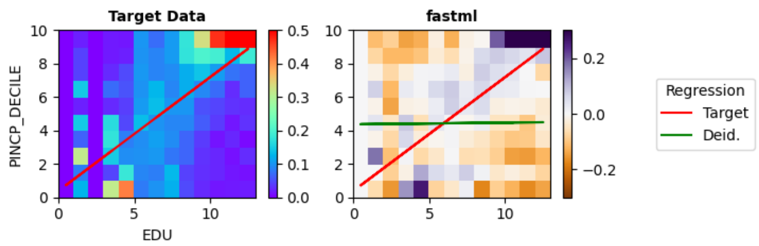
Feature Set: all-features | Target Dataset: national2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'DENSITY', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000



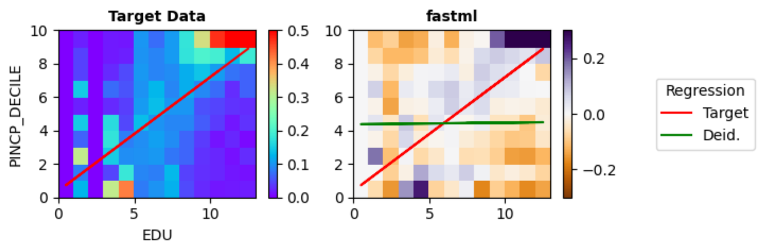
Feature Set: custom-features-23 | Target Dataset: national2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000



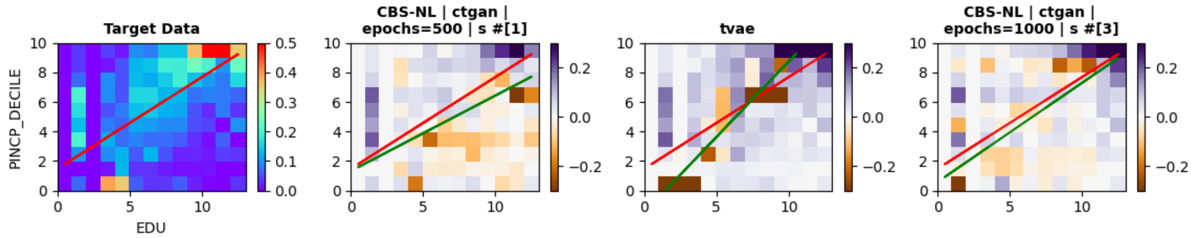
Feature Set: demographic_focused | Target Dataset: national2019:

Features: ['DEYE', 'SEX', 'HOUSING_TYPE', 'OWN_RENT', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'AGEP']
Feature Space (possible combinations): 227,026,800



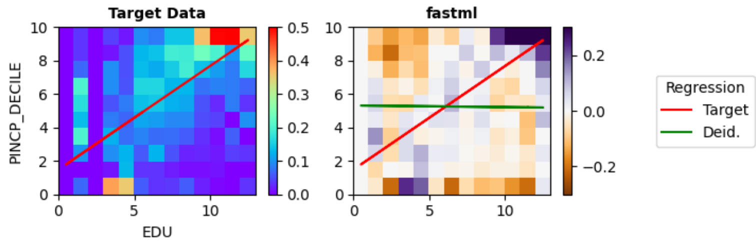
Feature Set: all-features | Target Dataset: tx2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'DENSITY', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000



Feature Set: custom-features-23 | Target Dataset: tx2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000



Regression Comparison: Black Women Data:

Linear regression is a fundamental data analysis technique that condenses a multi-dimensional data distribution down to a one dimensional (line) representation. It works by finding the line that sits in the 'middle' of the data, in some sense-- it minimizes the total distance between the points of the data and the line. There are more advanced forms of regression, but here we're focusing on the simplest case-- we fit a simple straight line to the data, getting the slope and y-intercept value of that line.

For this metric we're just looking at data from adults (AGEP > 15) and we're only considering the distribution of the data across two features:

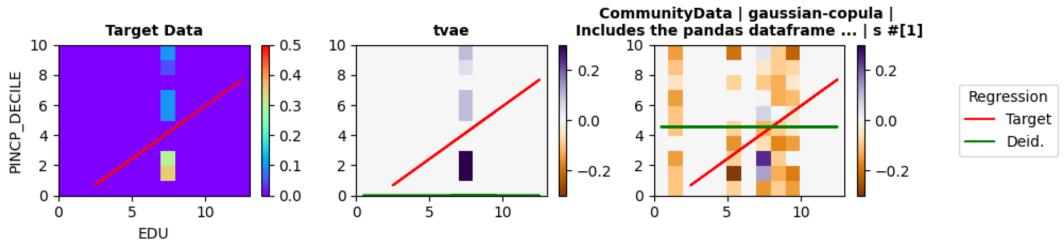
- EDU: The highest education level this individual has attained, ranging from 1 (elementary school) to 12 (PhD). See Appendix of this report for the full list of code values.
- PINCP_DECILE: The individual's income decile relative to their PUMA. This helps us account for differences in cost of living across the country. If an individual makes a moderate income but lives in a very low income area, they may have a high value for PINCP_DECILE indicating that they have a high income for their PUMA).

The basic idea is that higher values of EDU should lead to higher values of PINCP_DECILE, and this is broadly true. However, it is known that the relationship between EDU and PINCP_DECILE is different for different demographic subgroups. The heatmaps in the left column below show the density distribution of the true data for each subgroup, normalized by education category (so the density values in each column sum to 1; note that when a cell in the heatmap contains too few people (< 20), it is left blank; its not expected that the deidentified data will match the original distribution precisely). The regression line is drawn in red over the heatmap, so you can see the relationship between the target data distribution and its linear regression analysis. In the right column for each subgroup we show how the deidentified data's regression line compares to the target data's regression line, along with a heatmap of the density differences between the two distributions. Redder areas are where the deidentified data has created too many people, bluer areas are where it's created too few people.

We've broken this metric down into demographic subgroups so we can see not only how well the privacy techniques preserve the overall relationship between these features, but also whether they preserve how that overall relationship is built up from the different relationships that hold at each major demographic subgroup. It's important that deidentification techniques preserve these distinct subgroup patterns for analysis.

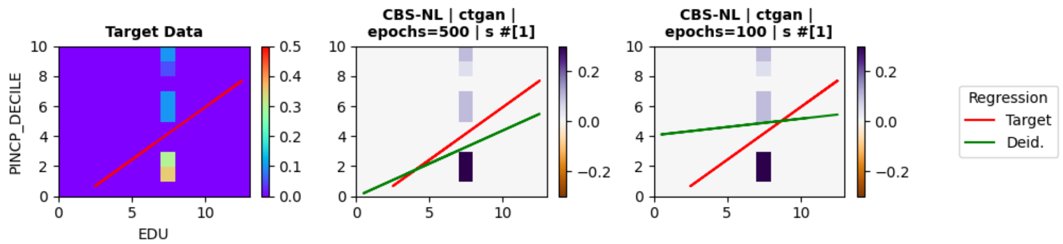
Feature Set: demographic-focused | Target Dataset: ma2019:

Features: ['DEYE', 'SEX', 'HOUSING_TYPE', 'OWN_RENT', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'AGEP']
Feature Space (possible combinations): 227,026,800



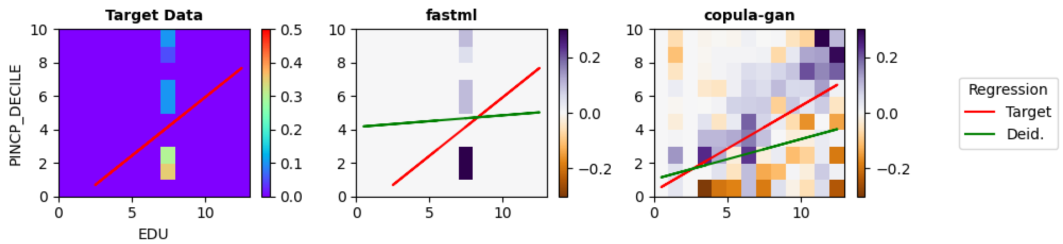
Feature Set: all-features | Target Dataset: ma2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'DENSITY', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000



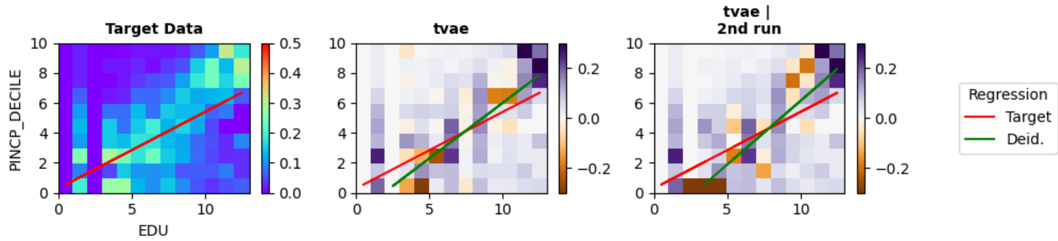
Feature Set: custom-features-23 | Target Dataset: ma2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000



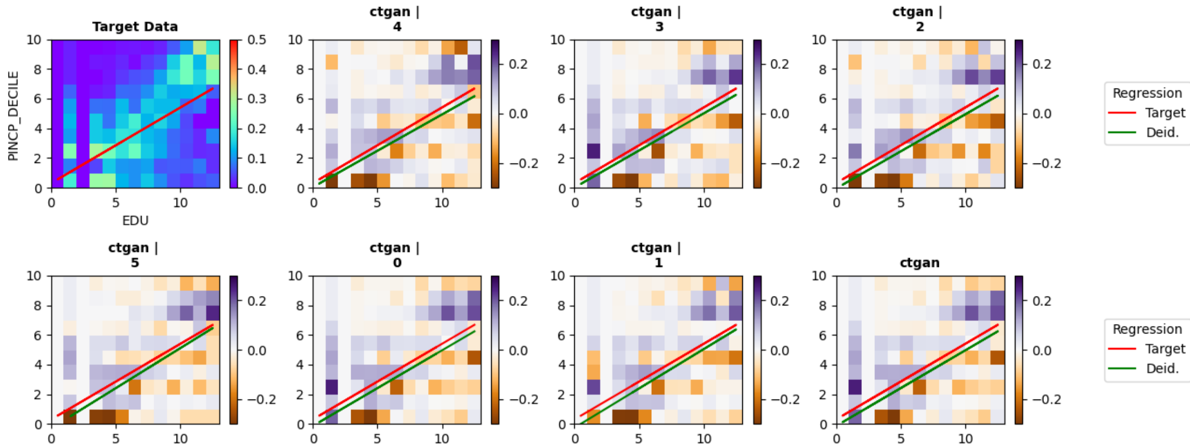
Feature Set: industry-focused | Target Dataset: national2019:

Features: ['SEX', 'OWN_RENT', 'HISP', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA']
Feature Space (possible combinations): 167,567,400



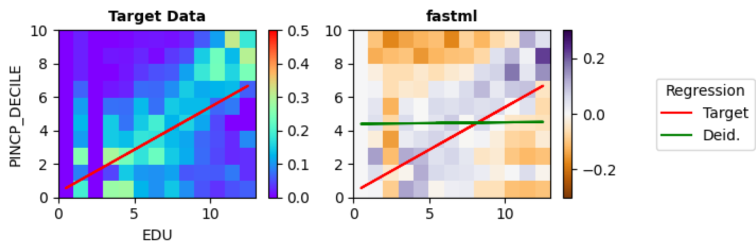
Feature Set: all-features | Target Dataset: national2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'DENSITY', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000



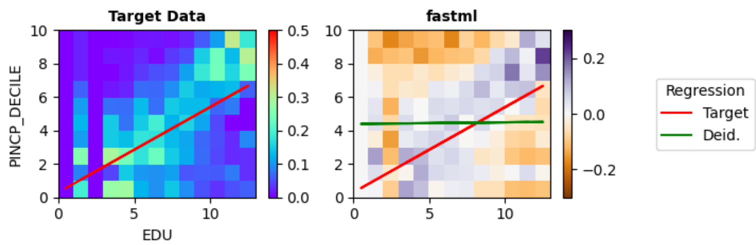
Feature Set: custom-features-23 | Target Dataset: national2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000



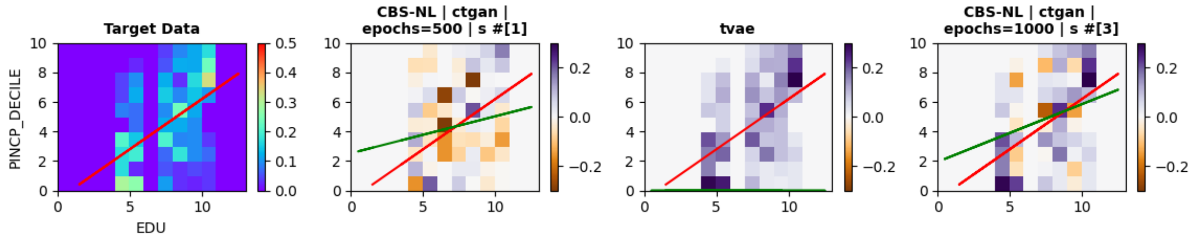
Feature Set: demographic_focused | Target Dataset: national2019:

Features: ['DEYE', 'SEX', 'HOUSING_TYPE', 'OWN_RENT', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'AGEP']
Feature Space (possible combinations): 227,026,800



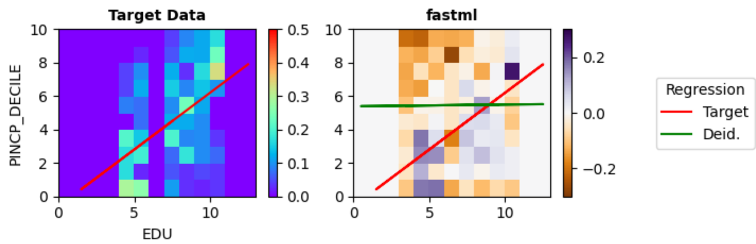
Feature Set: all-features | Target Dataset: tx2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'DENSITY', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000



Feature Set: custom-features-23 | Target Dataset: tx2019:

Features: ['SEX', 'DEAR', 'DEYE', 'OWN_RENT', 'HOUSING_TYPE', 'DPHY', 'DREM', 'HISP', 'DVET', 'MSP', 'RAC1P', 'PINCP_DECILE', 'EDU', 'INDP_CAT', 'PUMA', 'AGEP', 'INDP', 'PINCP', 'NOC', 'NPF', 'PWGTP', 'POVPIP', 'WGTP']
Feature Space (possible combinations): 3,407,717,673,360,000



Observations

The SDV library gives us a good opportunity to compare statistical modeling with neural network techniques. Both approaches have the possibility of introducing artifacts or bias as they work to capture the target data distribution. Looking at the PCA metric, how do the Gaussian Copulas differ from the GAN techniques? How do both differ from the TVAE technique?

The samples we have from SDV also explore a few other concepts. One is stability--we have many samples from the CTGAN approach. How do they compare with each other? Do different runs produce very different (unstable) or mostly similar (stable) results?

Another concept is tuning--the Gaussian Copula has two samples, one untuned (with default parameters) and on tuned (with parameters chosen to improve utility). What types of improvements did tuning make? What issues remained despite tuning?

Similarly, compare FastML with CTGAN; what types of improvements does the more advanced CTGAN method offer in comparison to the more primitive FastML? Increasing the training time (epochs) can improve performance for the CTGAN, but is it always true that more epochs means better utility? What structures in the target data pose difficulties for all GAN methods?

Data Description

Deidentified (Deid.) Datasets:

sdv | tvae | demographic-focused | :

Label Name	Label Value
Algorithm Name	tvae
Library	sdv
Feature Set	demographic-focused
Target Dataset	ma2019
Variant Label	

Property	Value
Filename	tvae_demographic
Records	7634
Features	10

sdv | gaussian-copula | demographic-focused | Includes the pandas dataframe index and an additional index_col that was need to get the synthesizer to run:

Label Name	Label Value
Team	CommunityData
Submission Timestamp	4/16/2023 20:31:04
Target Dataset	ma2019
Algorithm Name	gaussian-copula
Library	sdv
Feature Set	demographic-focused
Variant Label	Includes the pandas dataframe index and an additional index_col that was need to get the synthesizer to run
Submission Number	1
Privacy	Synthetic Data (Non-differentially Private)

Property	Value
Filename	sdv_gaussian_copula_JacobPasner_1
Records	7634
Features	10

sdv | tvae | industry-focused | :

Label Name	Label Value
Algorithm Name	tvae
Library	sdv
Feature Set	industry-focused
Target Dataset	national2019
Variant Label	

Property	Value
Filename	tvae_industry
Records	27253
Features	9

sdv | tvae | industry-focused | 2nd run:

Label Name	Label Value
Algorithm Name	tvae
Library	sdv
Feature Set	industry-focused
Target Dataset	national2019
Variant Label	2nd run

Property	Value
Filename	tvae_industry_2
Records	27254
Features	9

sdv | tvae | family-focused | :

Label Name	Label Value
Algorithm Name	tvae
Library	sdv
Feature Set	family-focused
Target Dataset	ma2019
Variant Label	

Property	Value
Filename	tvae_family
Records	7634
Features	11

sdv | gaussian-copula | family-focused | defaults:

Label Name	Label Value
Team	Blizzard Wizard
Submission Timestamp	4/11/2023 21:45:37
Target Dataset	national2019
Algorithm Name	gaussian-copula
Library	sdv
Feature Set	family-focused
Variant Label	defaults
Submission Number	1
Privacy	Synthetic Data (Non-differentially Private)

Property	Value
Filename	sdv_gaussian_copula_AaronMcMillin_1
Records	27253
Features	11

sdv | gaussian-copula | family-focused | tuned:

Label Name	Label Value
Team	Blizzard Wizard
Submission Timestamp	5/18/2023 21:45:37
Target Dataset	national2019
Algorithm Name	gaussian-copula
Library	sdv
Feature Set	family-focused
Variant Label	tuned
Submission Number	2
Privacy	Synthetic Data (Non-differentially Private)

Property	Value
Filename	sdv_gaussina_copula_tuned_family_AaronMcMillin_2
Records	27253
Features	11

sdv | ctgan | all-features | epochs=500:

Label Name	Label Value
Team	CBS-NL
Submission Timestamp	4/16/2023 12:03:58
Target Dataset	tx2019
Algorithm Name	ctgan
Library	sdv
Feature Set	all-features
Variant Label	epochs=500
Submission Number	1
Privacy	Synthetic Data (Non-differentially Private)

Property	Value
Filename	sdv_ctgan_epochs500-SlokomManel_1
Records	9276
Features	24

sdv | tvae | all-features | :

Label Name	Label Value
Algorithm Name	tvae
Library	sdv
Feature Set	all-features
Target Dataset	tx2019
Variant Label	

Property	Value
Filename	tvae_all
Records	9276
Features	24

sdv | ctgan | all-features | 4:

Label Name	Label Value
Algorithm Name	ctgan
Library	sdv
Feature Set	all-features
Target Dataset	national2019
Variant Label	4
Privacy	Synthetic Data (Non-differentially Private)

Property	Value
Filename	ctgan_4_AnnElliot
Records	27253
Features	24

sdv | ctgan | all-features | 3:

Label Name	Label Value
Algorithm Name	ctgan
Library	sdv
Feature Set	all-features
Target Dataset	national2019
Variant Label	3
Privacy	Synthetic Data (Non-differentially Private)

Property	Value
Filename	ctgan_3_AnnElliot
Records	27253
Features	24

sdv | ctgan | all-features | 2:

Label Name	Label Value
Algorithm Name	ctgan
Library	sdv
Feature Set	all-features
Target Dataset	national2019
Variant Label	2
Privacy	Synthetic Data (Non-differentially Private)

Property	Value
Filename	ctgan_2_AnnElliot
Records	27253
Features	24

sdv | ctgan | all-features | 5:

Label Name	Label Value
Algorithm Name	ctgan
Library	sdv
Feature Set	all-features
Target Dataset	national2019
Variant Label	5
Privacy	Synthetic Data (Non-differentially Private)

Property	Value
Filename	ctgan_5_AnnElliot
Records	27253
Features	24

sdv | ctgan | all-features | 0:

Label Name	Label Value
Algorithm Name	ctgan
Library	sdv
Feature Set	all-features
Target Dataset	national2019
Variant Label	0
Privacy	Synthetic Data (Non-differentially Private)

Property	Value
Filename	ctgan_0_AnnElliot
Records	27253
Features	24

sdv | ctgan | all-features | 1:

Label Name	Label Value
Algorithm Name	ctgan
Library	sdv
Feature Set	all-features
Target Dataset	national2019
Variant Label	1
Privacy	Synthetic Data (Non-differentially Private)

Property	Value
Filename	ctgan_1_AnnElliot
Records	27253
Features	24

sdv | ctgan | all-features | :

Label Name	Label Value
Algorithm Name	ctgan
Library	sdv
Feature Set	all-features
Target Dataset	national2019
Variant Label	
Privacy	Synthetic Data (Non-differentially Private)

Property	Value
Filename	ctgan_AnnElliot
Records	27253
Features	24

sdv | ctgan | all-features | epochs=500:

Label Name	Label Value
Team	CBS-NL
Submission Timestamp	4/16/2023 12:03:58
Target Dataset	ma2019
Algorithm Name	ctgan
Library	sdv
Feature Set	all-features
Variant Label	epochs=500
Submission Number	1
Privacy	Synthetic Data (Non-differentially Private)

Property	Value
Filename	sdv_ctgan_epochs500_SlokomManel_1
Records	7634
Features	24

sdv | ctgan | all-features | epochs=1000:

Label Name	Label Value
Team	CBS-NL
Submission Timestamp	4/28/2023 12:17:36
Target Dataset	tx2019
Algorithm Name	ctgan
Library	sdv
Feature Set	all-features
Variant Label	epochs=1000
Submission Number	3
Privacy	Synthetic Data (Non-differentially Private)

Property	Value
Filename	sdv_ctgan_epochs1000_SlokomManel_3
Records	9276
Features	24

sdv | ctgan | all-features | epochs=100:

Label Name	Label Value
Team	CBS-NL
Submission Timestamp	4/16/2023 12:03:58
Target Dataset	ma2019
Algorithm Name	ctgan
Library	sdv
Feature Set	all-features
Variant Label	epochs=100
Submission Number	1
Privacy	Synthetic Data (Non-differentially Private)

Property	Value
Filename	sdv_ctgan_epochs100_SlokomManel_1
Records	7634
Features	24

sdv | fastml | custom-features-23 | :

Label Name	Label Value
Algorithm Name	fastml
Library	sdv
Feature Set	custom-features-23
Target Dataset	ma2019
Variant Label	
Privacy	Synthetic Data (Non-differentially Private)

Property	Value
Filename	fastml_cf23_ma2019
Records	7634
Features	23

sdv | copula-gan | custom-features-23 | :

Label Name	Label Value
Algorithm Name	copula-gan
Library	sdv
Feature Set	custom-features-23
Target Dataset	ma2019
Variant Label	
Privacy	Synthetic Data (Non-differentially Private)

Property	Value
Filename	copulagan_cf23_na2019
Records	27253
Features	23

sdv | fastml | custom-features-23 | :

Label Name	Label Value
Algorithm Name	fastml
Library	sdv
Feature Set	custom-features-23
Target Dataset	national2019
Variant Label	
Privacy	Synthetic Data (Non-differentially Private)

Property	Value
Filename	fastml_cf23_na2019
Records	27253
Features	23

sdv | fastml | custom-features-23 | :

Label Name	Label Value
Algorithm Name	fastml
Library	sdv
Feature Set	custom-features-23
Target Dataset	tx2019
Variant Label	
Privacy	Synthetic Data (Non-differentially Private)

Property	Value
Filename	fastml_cf23_tx2019
Records	9276
Features	23

sdv | fastml | demographic_focused | :

Label Name	Label Value
Algorithm Name	fastml
Library	sdv
Feature Set	demographic_focused
Target Dataset	national2019
Variant Label	
Privacy	Synthetic Data (Non-differentially Private)

Property	Value
Filename	fastml_demographic_focused_na2019
Records	27253
Features	10

Appendix

Data Dictionary:

PUMA: Public use microdata area code:

PUMA Code	Code Description
25-00503	Middlesex County--Waltham City, Lexington, Burlington, Bedford & Lincoln Towns
25-00703	Essex County (East)--Salem, Beverly, Gloucester & Newburyport Cities
25-01000	Peabody City, Danvers, Reading, North Reading & Lynnfield Towns
25-01300	Billerica, Andover, Tewksbury & Wilmington Towns
25-02800	Woburn, Melrose Cities, Saugus, Wakefield & Stoneham Towns
48-02510	Tarrant County (North)--North Richland Hills (North) & Keller Cities
48-02102	Johnson County
48-02101	Ellis County
48-02515	Tarrant County (West)--Fort Worth City (West)
48-02507	Tarrant County (East)--Arlington City (West)--South of I-30 & East of Loop I-820
48-02516	Tarrant County (Southwest)--Fort Worth (Southwest) & Benbrook Cities
01-01301	Birmingham City (West)
06-07502	San Francisco County (North & East)--North Beach & Chinatown
06-08507	Santa Clara County (Southwest)--Cupertino, Saratoga Cities & Los Gatos Town
08-00803	Boulder County (Central)--Boulder City
13-04600	Atlanta Regional Commission--Fulton County (Central)--Atlanta City (Central)
17-03529	Chicago City (South)--South Shore, Hyde Park, Woodlawn, Grand Boulevard & Douglas
17-03531	Chicago City (South)--Auburn Gresham, Roseland, Chatham, Avalon Park & Burnside
19-01700	Des Moines City
24-01004	Montgomery County (South)--Bethesda, Potomac & North Bethesda
26-02702	Washtenaw County (East Central)--Ann Arbor City Area
28-01100	Central Region--Jackson City (East & Central)
29-01901	St. Louis City (North)
30-00600	East Montana (Outside Billings City)
32-00405	Las Vegas City (Southeast)
36-03710	NYC-Bronx Community District 1 & 2--Hunts Point, Longwood & Melrose
36-04010	NYC-Brooklyn Community District 17--East Flatbush, Farragut & Rugby
38-00100	West North Dakota--Minot City
40-00200	Cherokee, Sequoyah & Adair Counties
51-01301	Arlington County (North)
51-51255	Alexandria City

AGEP: Person's age:

AGEP Code	Code Description
min	0
max	99

SEX: Person's gender:

SEX Code	Code Description
1	Male
2	Female

MSP: Marital Status:

MSP Code	Code Description
N	N/A (age less than 15 years)
1	Now married, spouse present
2	Now Married, spouse absent
3	Widowed
4	Divorced
5	Separated
6	Never married

HISP: Hispanic origin:

HISP Code	Code Description
0	Not Spanish/Hispanic/Latino
1	Mexican
2	Puerto Rican
3	Cuban
4	All other Spanish/Hispanic/Latino

RAC1P: Person's Race:

RAC1P Code	Code Description
1	White alone
2	Black or African American alone
3	American Indian alone
4	Alaska Native alone
5	American Indian and Alaska Native tribes specified; or American Indian or Alaska Native, not specified and no other races
6	Asian alone
7	Native Hawaiian and Other Pacific Islander alone
8	Some Other Race alone
9	Two or More Races

NOC: Number of own children in household (unweighted):

NOC Code	Code Description
N	N/A (GQ/vacant)
0	No own children
min	1
max	19

NPF: Number of persons in family (unweighted):

NPF Code	Code Description
N	N/A (GQ/vacant/non-family household
min	2
max	20

HOUSING_TYPE: Housing unit or group quarters:

HOUSING_TYPE Code	Code Description
1	Housing Unit
2	Institutional Group Quarters
3	Non-institutional Group Quarters

OWN_RENT: Housing unit rented or owned:

OWN_RENT Code	Code Description
0	Group quarters
1	Own housing unit
2	Rent housing unit

DENSITY: Population density among residents of each PUMA:

DENSITY Code	Code Description
min	16.3
max	52864.7

Density Bin: 0 l Bin Range: (0, 150]

PUMA	DENSITY	PUMA NAME
30-00600	16.3	East Montana (Outside Billings City)
38-00100	73.0	West North Dakota--Minot City
40-00200	90.7	Cherokee, Sequoyah & Adair Counties

Density Bin: 2 l Bin Range: (309.67, 475.62]

PUMA	DENSITY	PUMA NAME
48-02101	357.4	Ellis County
48-02102	450.9	Johnson County

Density Bin: 5 l Bin Range: (1121.99, 1723.27]

PUMA	DENSITY	PUMA NAME
25-01300	1457.2	Billerica, Andover, Tewksbury & Wilmington Towns
48-02516	1338.4	Tarrant County (Southwest)--Fort Worth (Southwest) & Benbrook Cities

Density Bin: 6 l Bin Range: (1723.27, 2646.76]

PUMA	DENSITY	PUMA NAME
25-00703	2195.3	Essex County (East)--Salem, Beverly, Gloucester & Newburyport Cities
25-01000	2447.1	Peabody City, Danvers, Reading, North Reading & Lynnfield Towns
48-02515	2134.8	Tarrant County (West)--Fort Worth City (West)

Density Bin: 7 l Bin Range: (2646.76, 4065.16]

PUMA	DENSITY	PUMA NAME
01-01301	2731.2	Birmingham City (West)
06-08507	3305.1	Santa Clara County (Southwest)--Cupertino, Saratoga Cities & Los Gatos Town
08-00803	3393.2	Boulder County (Central)--Boulder City
13-04600	3670.4	Atlanta Regional Commission--Fulton County (Central)--Atlanta City (Central)
19-01700	3572.3	Des Moines City
25-00503	2872.7	Middlesex County--Waltham City, Lexington, Burlington, Bedford & Lincoln Towns
25-02800	3683.9	Woburn, Melrose Cities, Saugus, Wakefield & Stoneham Towns
28-01100	2674.3	Central Region--Jackson City (East & Central)
48-02507	3731.1	Tarrant County (East)--Arlington City (West)--South of I-30 & East of Loop I-820
48-02510	3092.4	Tarrant County (North)--North Richland Hills (North) & Keller Cities

Density Bin: 8 l Bin Range: (4065.16, 6243.68]

PUMA	DENSITY	PUMA NAME
24-01004	4187.9	Montgomery County (South)--Bethesda, Potomac & North Bethesda
26-02702	4817.2	Washtenaw County (East Central)--Ann Arbor City Area
29-01901	5434.8	St. Louis City (North)

Density Bin: 9 l Bin Range: (6243.68, 9589.66]

PUMA	DENSITY	PUMA NAME
32-00405	7990.5	Las Vegas City (Southeast)

Density Bin: 10 l Bin Range: (9589.66, 14728.75]

PUMA	DENSITY	PUMA NAME
17-03531	11171.6	Chicago City (South)--Auburn Gresham, Roseland, Chatham, Avalon Park & Burnside
51-01301	11162.8	Arlington County (North)
51-51255	11224.3	Alexandria City

Density Bin: 11 l Bin Range: (14728.75, 22621.88]

PUMA	DENSITY	PUMA NAME
17-03529	15097.5	Chicago City (South)--South Shore, Hyde Park, Woodlawn, Grand Boulevard & Douglas

Density Bin: 12 l Bin Range: (22621.88, 34744.92]

PUMA	DENSITY	PUMA NAME
06-07502	33632.6	San Francisco County (North & East)--North Beach & Chinatown

Density Bin: 13 l Bin Range: (34744.92, 53364.7]

PUMA	DENSITY	PUMA NAME
36-03710	52864.7	NYC-Bronx Community District 1 & 2--Hunts Point, Longwood & Melrose
36-04010	50441.6	NYC-Brooklyn Community District 17--East Flatbush, Farragut & Rugby

INDP: Industry codes:

See codes in ACS data dictionary. Find codes by searching the string: INDP, in the ACS data dictionary

INDP_CAT: Industry categories:

INDP_CAT Code	Code Description
N	N/A (less than 16 years old, or last worked more than 5 years ago, or never worked)
0	AGR: Agriculture, Forestry, Fishing and Hunting
1	EXT: Mining, Quarrying, and Oil and Gas Extraction
2	UTL: Utilities
3	CON: Construction
4	MFG: Manufacturing
5	WHL: Wholesale Trade
6	RET: Retail Trade
7	TRN: Transportation and Warehousing
8	INF: Information
9	FIN: Finance, Insurance, Real Estate
10	PRF: Professional, Scientific and Technical Services
11	EDU: Educational Services
12	MED: Health Care
13	SCA: Social Assistance
14	ENT: Arts, Entertainment, Accommodation, Food Services and Recreation
15	SRV: Other Services
16	ADM: Government, Public Administration
17	MIL: Military
18	UNEMPLOYED

EDU: Educational attainment:

EDU Code	Code Description
N	N/A (less than 3 years old)
1	No schooling completed
2	Nursery school, Preschool, or Kindergarten
3	Grade 1 to grade 8
4	Grade 9 to grade 12, no diploma
5	High School diploma
6	GED
7	Some College, no degree
8	Associate degree
9	Bachelors degree
10	Masters degree
11	Professional degree
12	Doctorate degree

PINCP: Person's total income in dollars:

PINCP Code	Code Description
N	N/A (less than 15 years old)
min	-9000
max	1341000

PINCP_DECILE: Person's total income rank (with respect to their state) discretized into 10% bins.:

PINCP_DECILE Code	Code Description
N	N/A (less than 15 years old
9	90th percentile
8	80th percentile
7	70th percentile
6	60th percentile
5	50th percentile
4	40th percentile
3	30th percentile
2	20th percentile
1	10th percentile
0	0th percentile

POVPIP: Income-to-poverty ratio (ex: 250 = 2.5 x poverty line):

POVPIP Code	Code Description
N	N/A
min	0
max	500
501	income above 5 x poverty line

DVET: Veteran service connected disability rating (percentage):

DVET Code	Code Description
N	N/A (No service-connected disability/never served in military
1	0 percent
2	10 or 20 percent
3	30 or 40 percent
4	50 or 60 percent
5	70, 80, 90 or 100 percent
6	Not reported

DREM: Cognitive difficulty:

DREM Code	Code Description
N	N/A (Less than 5 years old)
1	Yes
2	No

DPHY: Ambulatory (walking) difficulty:

DPHY Code	Code Description
N	N/A (Less than 5 years old)
1	Yes
2	No

DEYE: Vision difficulty:

DEYE Code	Code Description
1	Yes
2	No

DEAR: Hearing difficulty:

DEAR Code	Code Description
1	Yes
2	No

WGTP: Housing unit sampling weight:

See description of weights.

WGTP Code	Code Description
0	Group quarters place holder record
min	1
max	9999

PWGTP: Person's sampling weight:

See description of weights.

PWGTP Code	Code Description
min	1
max	9999