NIST CRC Meta Report
Subsample

Report created on: January 31, 2024 21:25:06

Created with SDNIST v2.3.0

Motivation

Arguably the simplest method used for Statistical Disclosure Control deidentification is just subsampling--a percentage of the data is randomly withheld from the release. This protects privacy by potentially enabling individuals who were known to have taken the survey to claim that their records were not included in the released data. The American Community Survey (from which we draw our target data excerpts that we're using in this project) uses subsampling as one step of their deidentification approach. According to their website, they collect about a 2.5% sample of the US population and release about 40% of that data (a 1% sample of the US population).

Subsampling has two different types of impact on the data, with respect to utility/equity and deidentificatio/privacy. For equity, especially on larger feature sets, subsampling can have disparate negative impact on smaller subpopulations. For privacy there's the caveat that real individuals' records are still being publicly released (see the scores for the unique exact match metric).

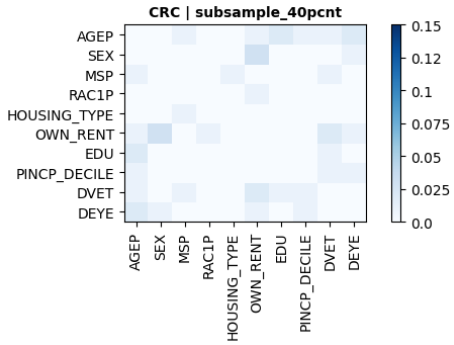
Comparisons

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.

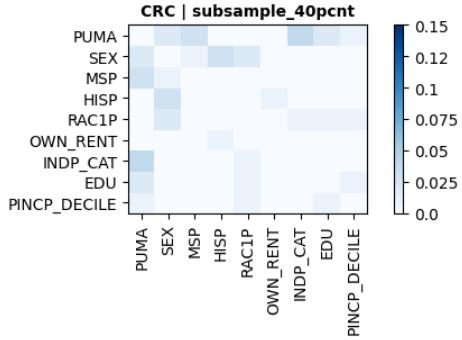
Feature Set: demographic-focused | Target Dataset: tx2019:

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



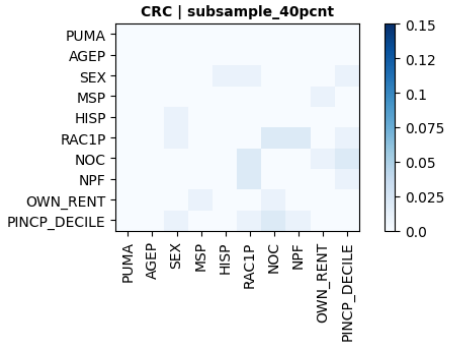
Feature Set: industry-focused | Target Dataset: tx2019:

Features: ['EDU', 'HISP', 'INDP_CAT', 'MSP', 'OWN_RENT', 'PINCP_DECILE', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 32,432,400



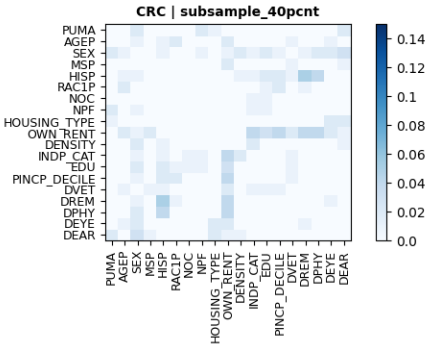
Feature Set: family-focused | Target Dataset: tx2019:

Features: ['AGEP', 'HISP', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 124,740,000,000



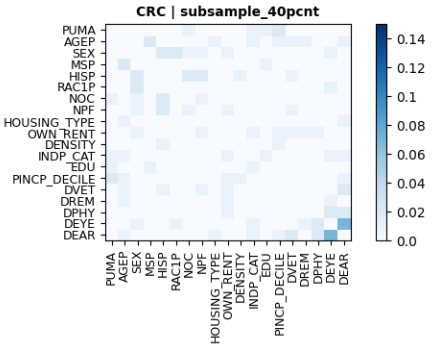
Feature Set: all-features | Target Dataset: tx2019:

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'PWGTP', 'RAC1P', 'SEX', 'WGTP']
Feature Space (possible combinations): 36,042,774,768,000,000,000,000,000



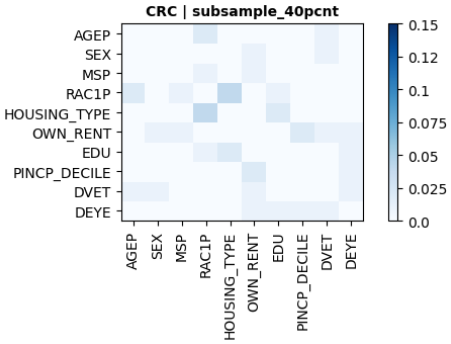
Feature Set: simple-features | Target Dataset: tx2019:

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 14,711,336,640,000,000,000



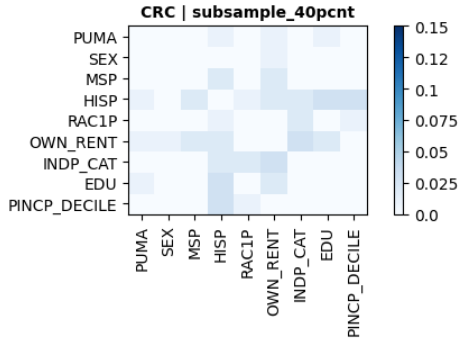
Feature Set: demographic-focused | Target Dataset: ma2019:

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



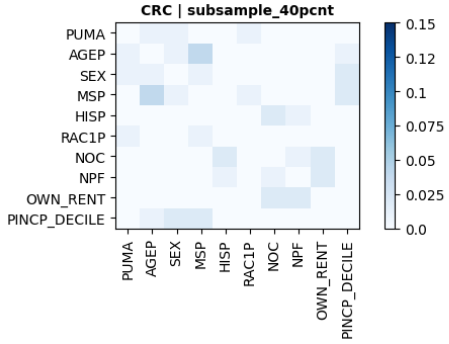
Feature Set: industry-focused | Target Dataset: ma2019:

Features: ['EDU', 'HISP', 'INDP_CAT', 'MSP', 'OWN_RENT', 'PINCP_DECILE', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 32,432,400



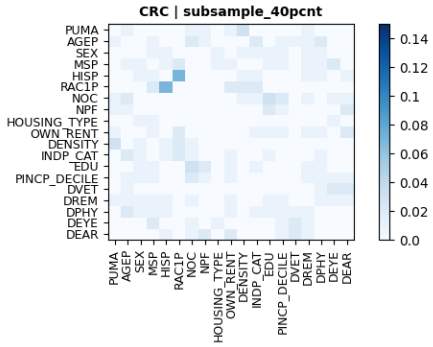
Feature Set: family-focused | Target Dataset: ma2019:

Features: ['AGEP', 'HISP', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 124,740,000,000



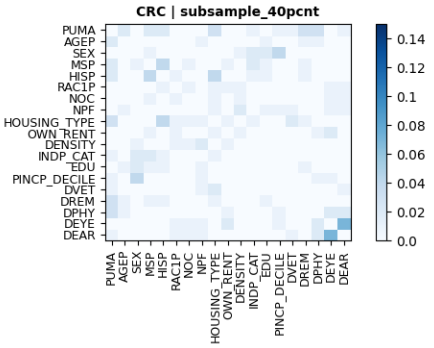
Feature Set: all-features | Target Dataset: ma2019:

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'PWGTP', 'RAC1P', 'SEX', 'WGTP']
Feature Space (possible combinations): 36,042,774,768,000,000,000,000,000



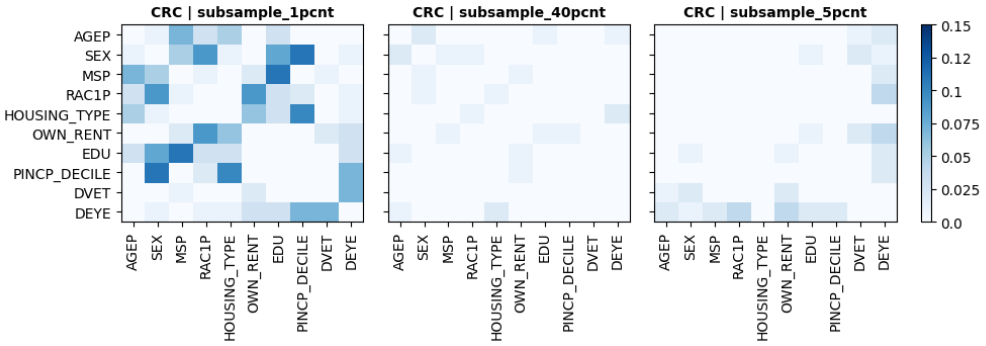
Feature Set: simple-features | Target Dataset: ma2019:

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 14,711,336,640,000,000,000



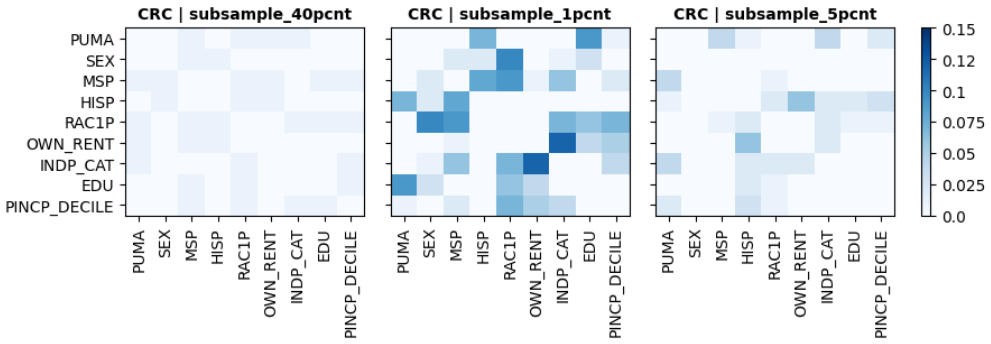
Feature Set: demographic-focused | Target Dataset: national2019:

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



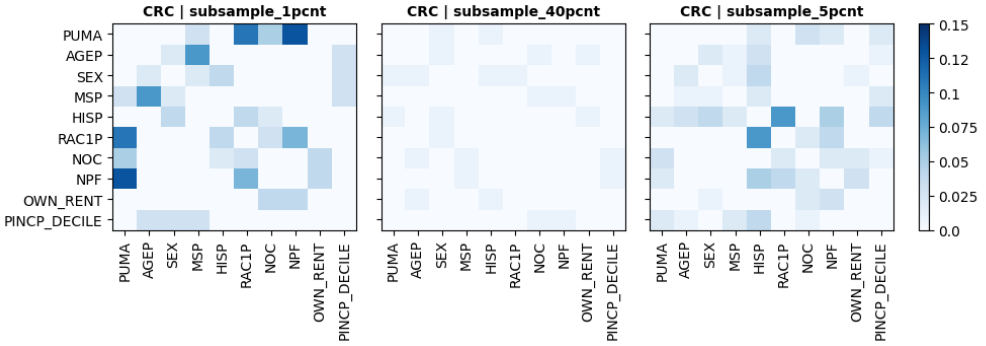
Feature Set: industry-focused | Target Dataset: national2019:

Features: ['EDU', 'HISP', 'INDP_CAT', 'MSP', 'OWN_RENT', 'PINCP_DECILE', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 32,432,400



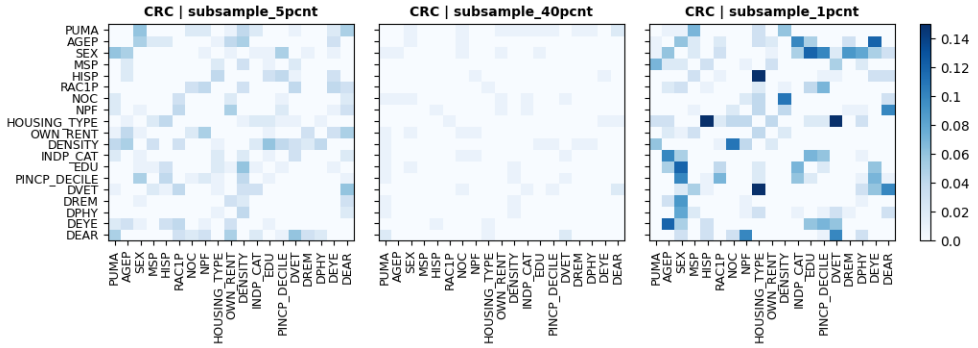
Feature Set: family-focused | Target Dataset: national2019:

Features: ['AGEP', 'HISP', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 124,740,000,000



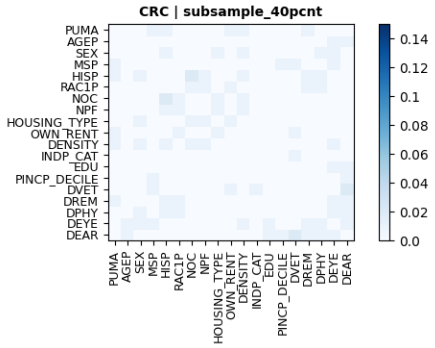
Feature Set: all-features | Target Dataset: national2019:

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'PWGTP', 'RAC1P', 'SEX', 'WGTP']
Feature Space (possible combinations): 36,042,774,768,000,000,000,000,000



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

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 14,711,336,640,000,000,000



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: tx2019:

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

Number of Unique Records in Target Data: 5268 (56.79%)

Variant	Records Matched In Target Data	Percent Records Matched In Target Data
CRC | subsample_40pcnt	2090	39.67

Feature Set: industry-focused | Target Dataset: tx2019:

Features: ['EDU', 'HISP', 'INDP_CAT', 'MSP', 'OWN_RENT', 'PINCP_DECILE', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 32,432,400

Number of Unique Records in Target Data: 5504 (59.34%)

Variant	Records Matched In Target Data	Percent Records Matched In Target Data
CRC | subsample_40pcnt	2204	40.04

Feature Set: family-focused | Target Dataset: tx2019:

Features: ['AGEP', 'HISP', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 124,740,000,000

Number of Unique Records in Target Data: 8601 (92.72%)

Variant	Records Matched In Target Data	Percent Records Matched In Target Data
CRC | subsample_40pcnt	3438	39.97

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

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'PWGTP', 'RAC1P', 'SEX', 'WGTP']
Feature Space (possible combinations): 36,042,774,768,000,000,000,000,000

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

Variant	Records Matched In Target Data	Percent Records Matched In Target Data
CRC | subsample_40pcnt	3707	40.03

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

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 14,711,336,640,000,000,000

Number of Unique Records in Target Data: 9075 (97.83%)

Variant	Records Matched In Target Data	Percent Records Matched In Target Data
CRC | subsample_40pcnt	3632	40.02

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

Features: ['AGEP', 'DEYE', 'DVET', 'EDU', 'HOUSING_TYPE', 'MSP', 'OWN_RENT', 'PINCP_DECILE', 'RAC1P', 'SEX']
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
CRC | subsample_40pcnt	1710	40.07

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

Features: ['EDU', 'HISP', 'INDP_CAT', 'MSP', 'OWN_RENT', 'PINCP_DECILE', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 32,432,400

Number of Unique Records in Target Data: 4265 (55.87%)

Variant	Records Matched In Target Data	Percent Records Matched In Target Data
CRC | subsample_40pcnt	1689	39.6

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

Features: ['AGEP', 'HISP', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 124,740,000,000

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

Variant	Records Matched In Target Data	Percent Records Matched In Target Data
CRC | subsample_40pcnt	2563	39.83

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

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'PWGTP', 'RAC1P', 'SEX', 'WGTP']
Feature Space (possible combinations): 36,042,774,768,000,000,000,000,000

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

Variant	Records Matched In Target Data	Percent Records Matched In Target Data
CRC | subsample_40pcnt	3051	40.01

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

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 14,711,336,640,000,000,000

Number of Unique Records in Target Data: 7176 (94.0%)

Variant	Records Matched In Target Data	Percent Records Matched In Target Data
CRC | subsample_40pcnt	2870	39.99

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

Features: ['AGEP', 'DEYE', 'DVET', 'EDU', 'HOUSING_TYPE', 'MSP', 'OWN_RENT', 'PINCP_DECILE', 'RAC1P', 'SEX']
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
CRC | subsample_1pcnt	142	0.95
CRC | subsample_40pcnt	5963	39.97
CRC | subsample_5pcnt	746	5.0

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

Features: ['EDU', 'HISP', 'INDP_CAT', 'MSP', 'OWN_RENT', 'PINCP_DECILE', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 32,432,400

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

Variant	Records Matched In Target Data	Percent Records Matched In Target Data
CRC | subsample_40pcnt	6382	39.56
CRC | subsample_1pcnt	161	1.0
CRC | subsample_5pcnt	792	4.91

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

Features: ['AGEP', 'HISP', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 124,740,000,000

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

Variant	Records Matched In Target Data	Percent Records Matched In Target Data
CRC | subsample_1pcnt	235	0.98
CRC | subsample_40pcnt	9578	40.06
CRC | subsample_5pcnt	1195	5.0

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

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'PWGTP', 'RAC1P', 'SEX', 'WGTP']
Feature Space (possible combinations): 36,042,774,768,000,000,000,000,000

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

Variant	Records Matched In Target Data	Percent Records Matched In Target Data
CRC | subsample_5pcnt	1354	4.99
CRC | subsample_40pcnt	10860	39.99
CRC | subsample_1pcnt	273	1.01

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

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 14,711,336,640,000,000,000

Number of Unique Records in Target Data: 25762 (94.53%)

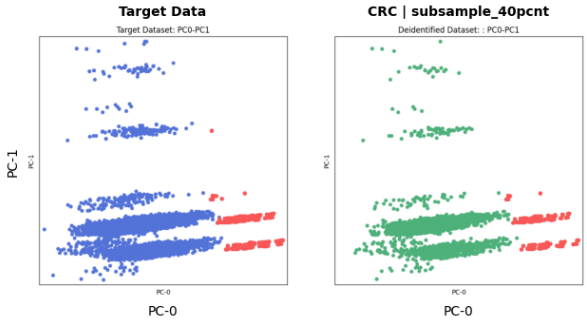
Variant	Records Matched In Target Data	Percent Records Matched In Target Data
CRC | subsample_40pcnt	10327	40.09

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.

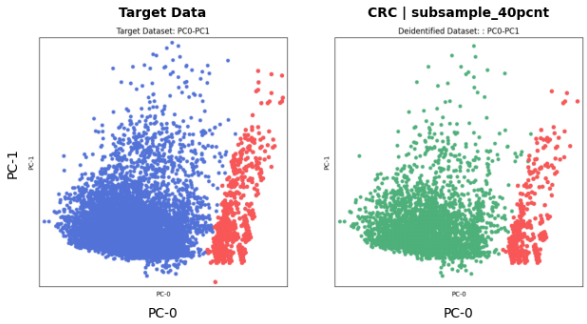
Feature Set: demographic-focused | Target Dataset: tx2019:

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



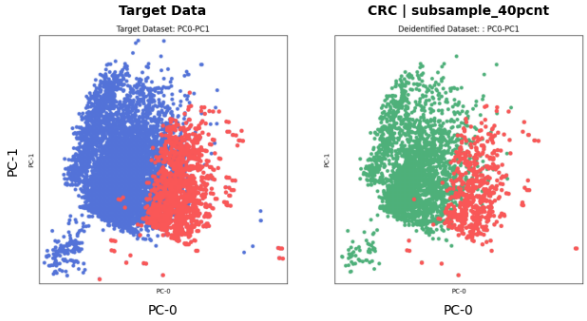
Feature Set: industry-focused | Target Dataset: tx2019:

Features: ['EDU', 'HISP', 'INDP_CAT', 'MSP', 'OWN_RENT', 'PINCP_DECILE', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 32,432,400



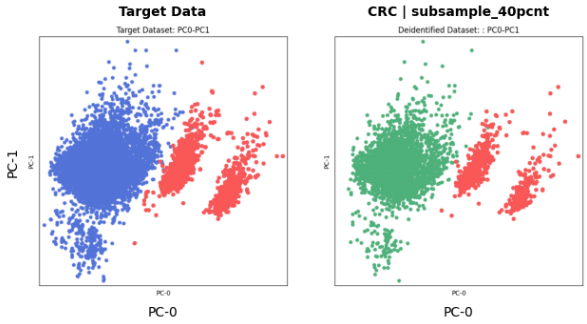
Feature Set: family-focused | Target Dataset: tx2019:

Features: ['AGEP', 'HISP', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 124,740,000,000



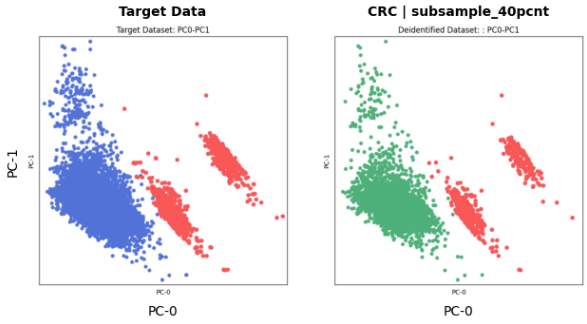
Feature Set: all-features | Target Dataset: tx2019:

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'PWGTP', 'RAC1P', 'SEX', 'WGTP']
Feature Space (possible combinations): 36,042,774,768,000,000,000,000,000



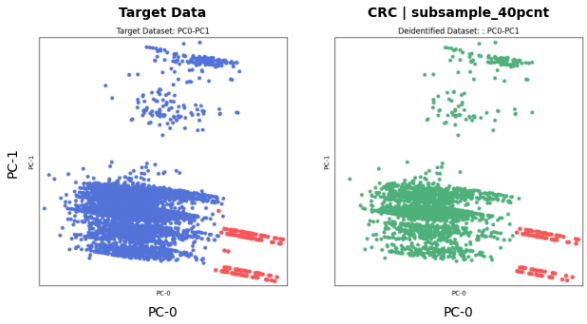
Feature Set: simple-features | Target Dataset: tx2019:

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 14,711,336,640,000,000,000



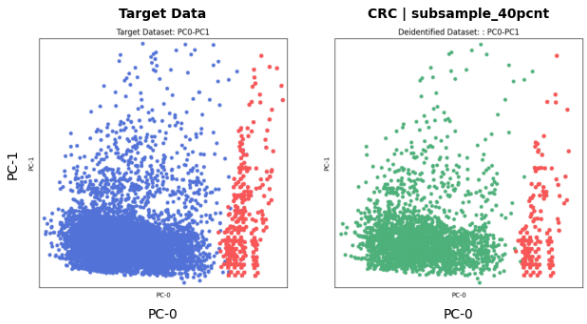
Feature Set: demographic-focused | Target Dataset: ma2019:

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



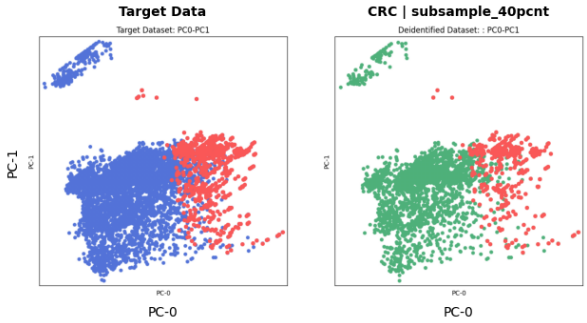
Feature Set: industry-focused | Target Dataset: ma2019:

Features: ['EDU', 'HISP', 'INDP_CAT', 'MSP', 'OWN_RENT', 'PINCP_DECILE', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 32,432,400



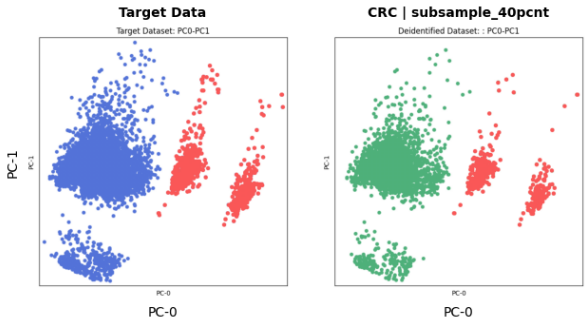
Feature Set: family-focused | Target Dataset: ma2019:

Features: ['AGEP', 'HISP', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 124,740,000,000



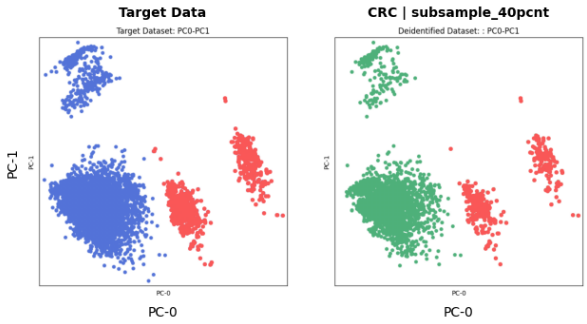
Feature Set: all-features | Target Dataset: ma2019:

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'PWGTP', 'RAC1P', 'SEX', 'WGTP']
Feature Space (possible combinations): 36,042,774,768,000,000,000,000,000



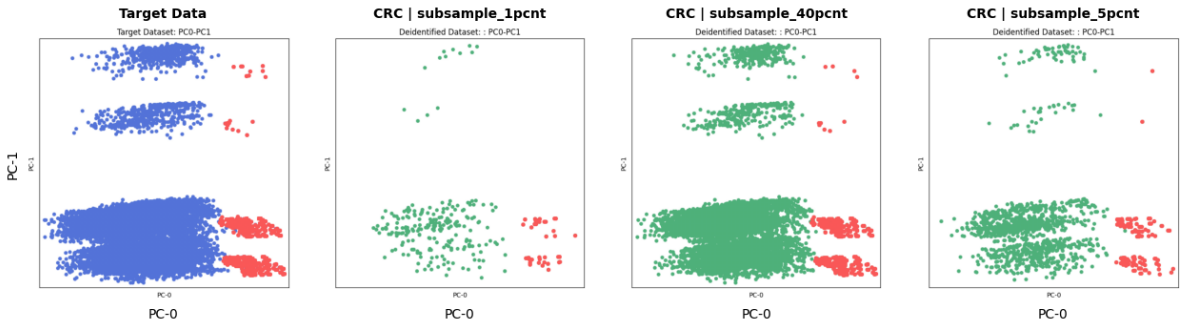
Feature Set: simple-features | Target Dataset: ma2019:

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 14,711,336,640,000,000,000



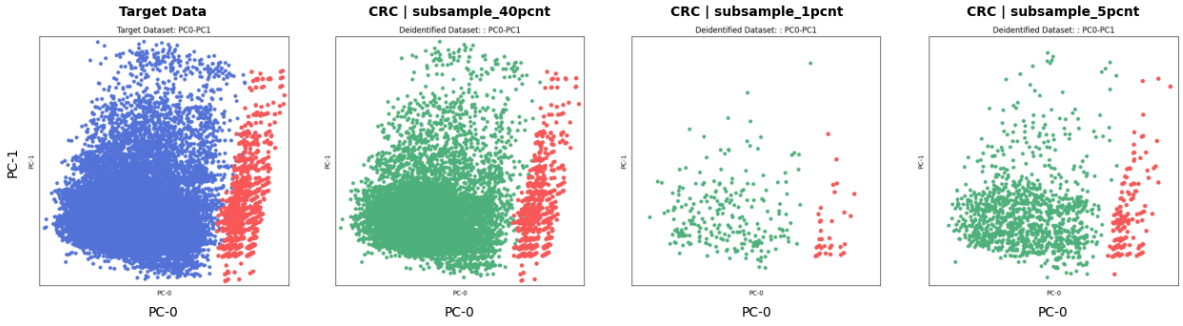
Feature Set: demographic-focused | Target Dataset: national2019:

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



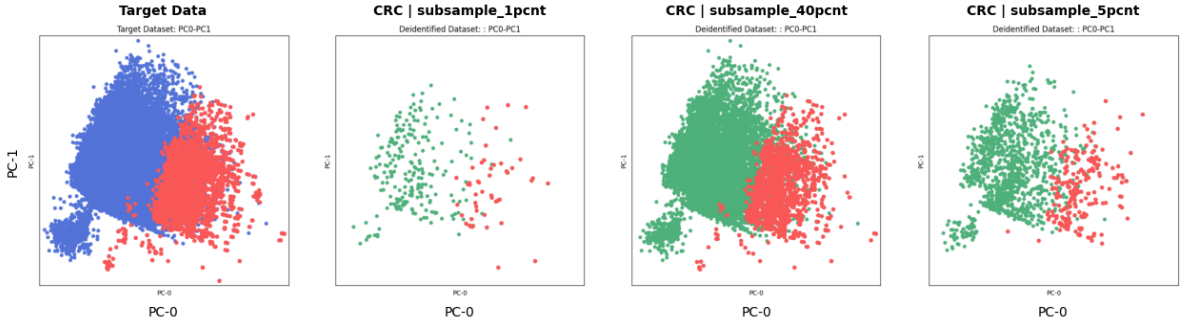
Feature Set: industry-focused | Target Dataset: national2019:

Features: ['EDU', 'HISP', 'INDP_CAT', 'MSP', 'OWN_RENT', 'PINCP_DECILE', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 32,432,400



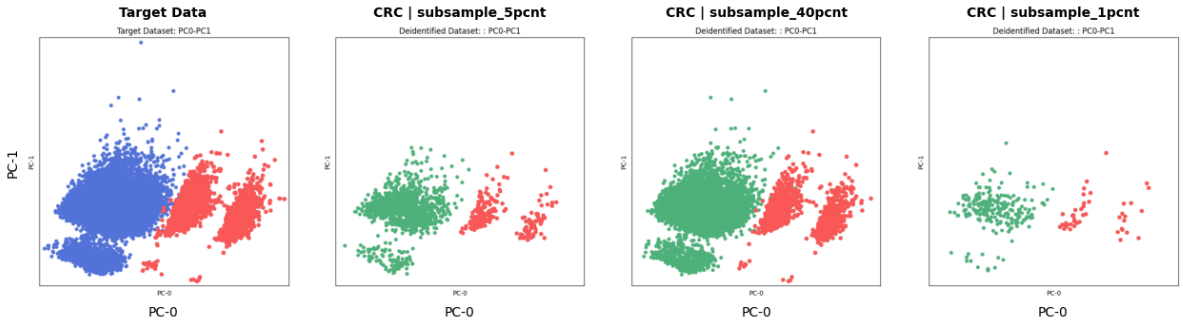
Feature Set: family-focused | Target Dataset: national2019:

Features: ['AGEP', 'HISP', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 124,740,000,000



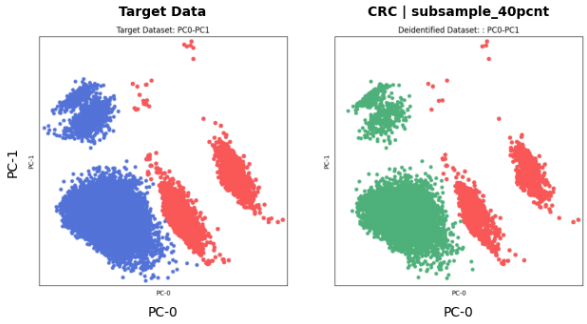
Feature Set: all-features | Target Dataset: national2019:

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'PWGTP', 'RAC1P', 'SEX', 'WGTP']
Feature Space (possible combinations): 36,042,774,768,000,000,000,000,000



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

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 14,711,336,640,000,000,000



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:

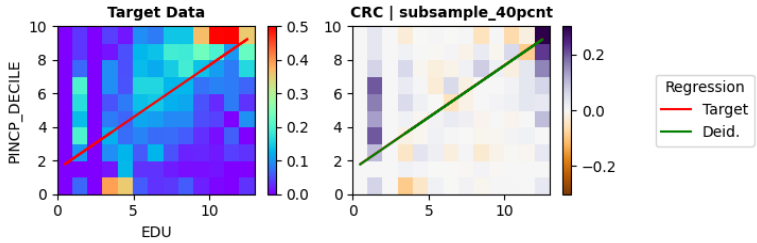
- EDU: The highest education level this individual has attained, ranging from 1 (elementary school) to 12 (PhD). See Appendix of this report for the full list of code values.
- PINCP_DECILE: The individual's income decile relative to their PUMA. This helps us account for differences in cost of living 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.

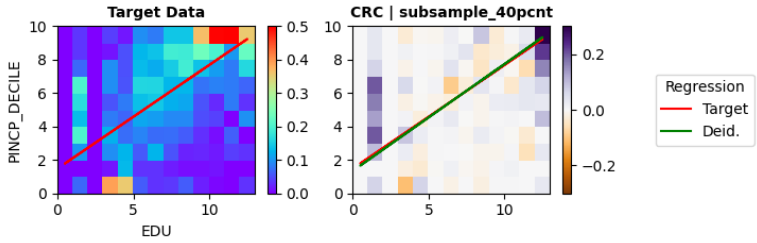
Feature Set: demographic-focused | Target Dataset: tx2019:

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



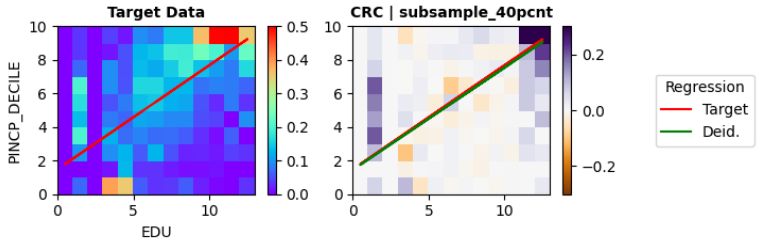
Feature Set: industry-focused | Target Dataset: tx2019:

Features: ['EDU', 'HISP', 'INDP_CAT', 'MSP', 'OWN_RENT', 'PINCP_DECILE', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 32,432,400



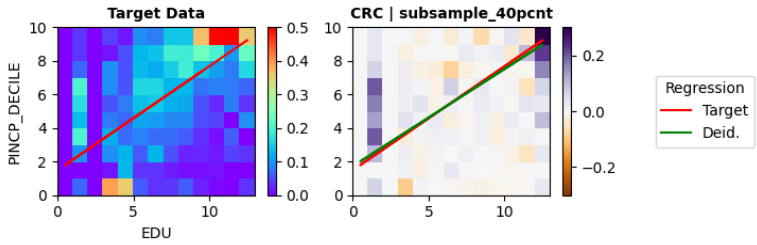
Feature Set: all-features | Target Dataset: tx2019:

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'PWGTP', 'RAC1P', 'SEX', 'WGTP']
Feature Space (possible combinations): 36,042,774,768,000,000,000,000,000



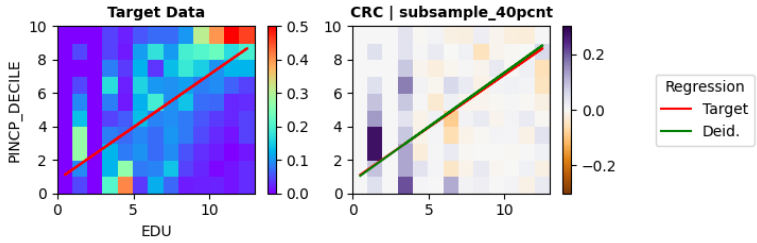
Feature Set: simple-features | Target Dataset: tx2019:

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 14,711,336,640,000,000,000



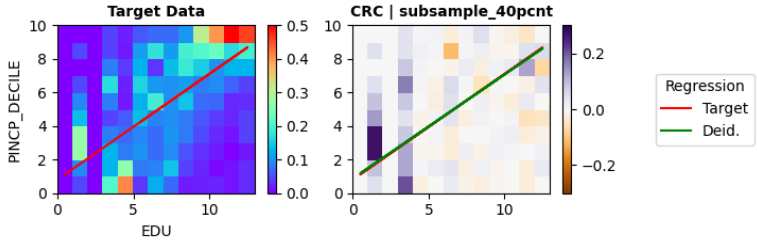
Feature Set: demographic-focused | Target Dataset: ma2019:

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



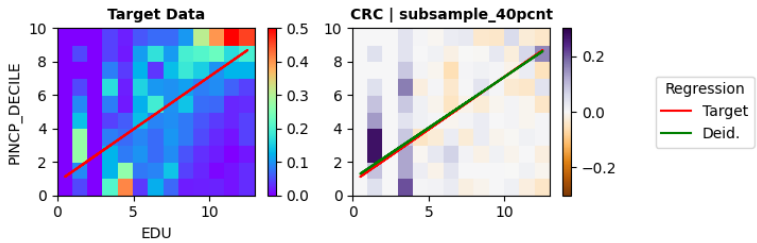
Feature Set: industry-focused | Target Dataset: ma2019:

Features: ['EDU', 'HISP', 'INDP_CAT', 'MSP', 'OWN_RENT', 'PINCP_DECILE', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 32,432,400



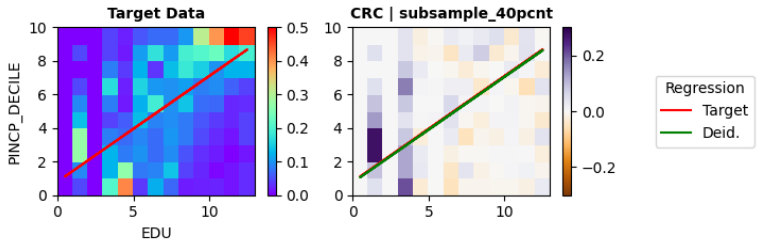
Feature Set: all-features | Target Dataset: ma2019:

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'PWGTP', 'RAC1P', 'SEX', 'WGTP']
Feature Space (possible combinations): 36,042,774,768,000,000,000,000,000



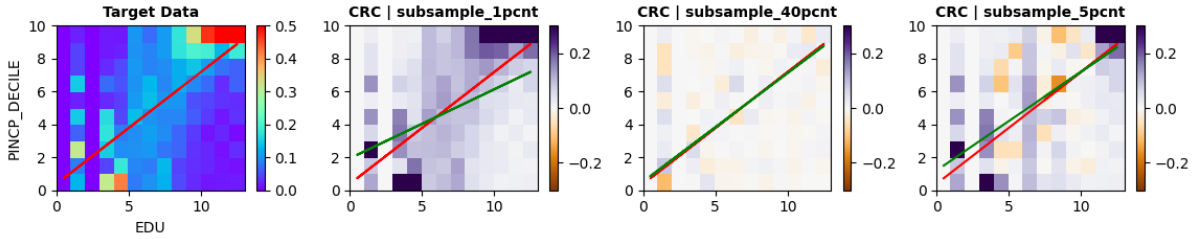
Feature Set: simple-features | Target Dataset: ma2019:

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 14,711,336,640,000,000,000



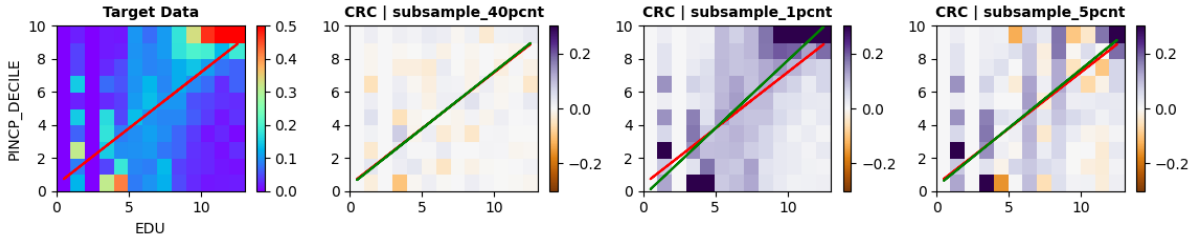
Feature Set: demographic-focused | Target Dataset: national2019:

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



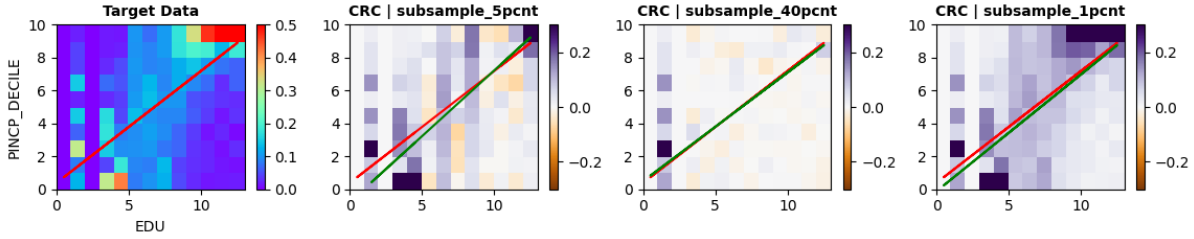
Feature Set: industry-focused | Target Dataset: national2019:

Features: ['EDU', 'HISP', 'INDP_CAT', 'MSP', 'OWN_RENT', 'PINCP_DECILE', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 32,432,400



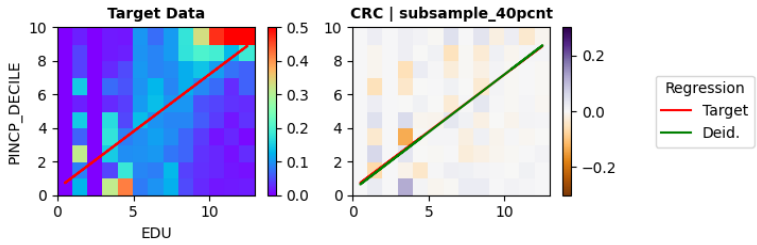
Feature Set: all-features | Target Dataset: national2019:

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'PWGTP', 'RAC1P', 'SEX', 'WGTP']
Feature Space (possible combinations): 36,042,774,768,000,000,000,000,000



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

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 14,711,336,640,000,000,000



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:

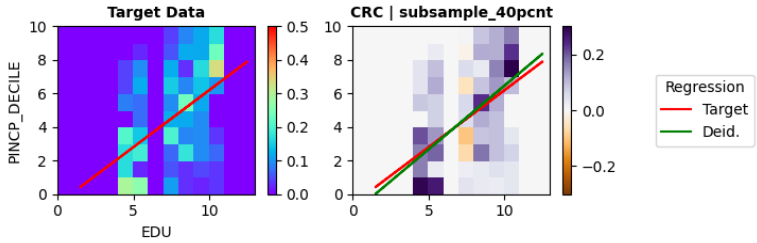
- EDU: The highest education level this individual has attained, ranging from 1 (elementary school) to 12 (PhD). See Appendix of this report for the full list of code values.
- PINCP_DECILE: The individual's income decile relative to their PUMA. This helps us account for differences in cost of living 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.

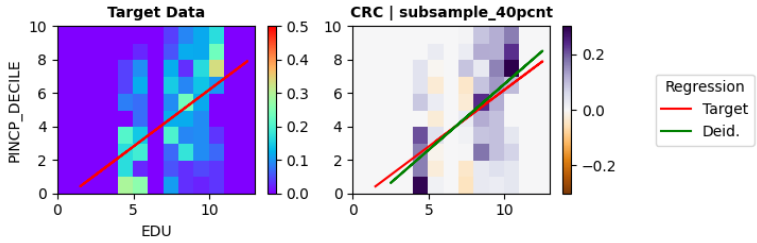
Feature Set: demographic-focused | Target Dataset: tx2019:

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



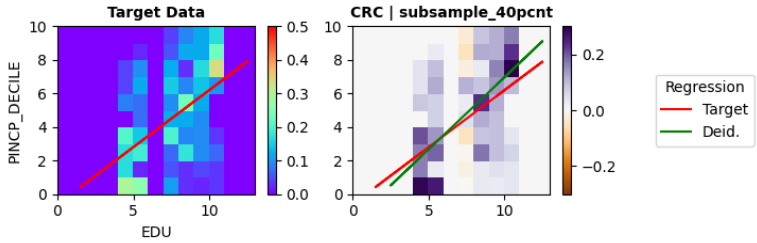
Feature Set: industry-focused | Target Dataset: tx2019:

Features: ['EDU', 'HISP', 'INDP_CAT', 'MSP', 'OWN_RENT', 'PINCP_DECILE', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 32,432,400



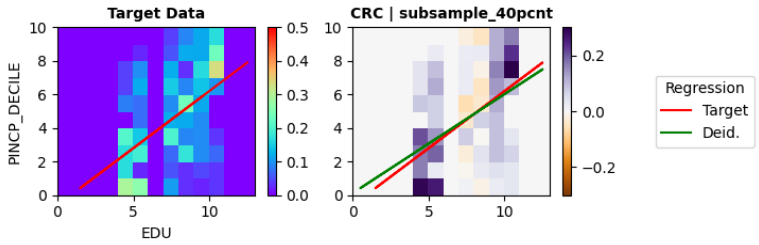
Feature Set: all-features | Target Dataset: tx2019:

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'PWGTP', 'RAC1P', 'SEX', 'WGTP']
Feature Space (possible combinations): 36,042,774,768,000,000,000,000,000



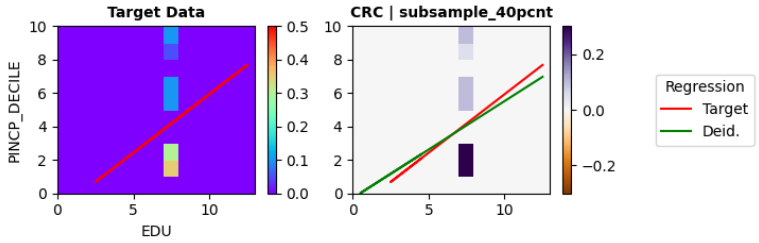
Feature Set: simple-features | Target Dataset: tx2019:

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 14,711,336,640,000,000,000



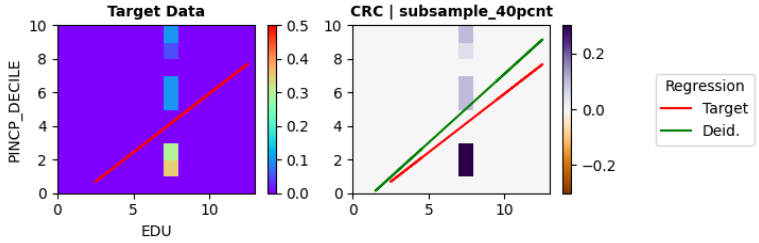
Feature Set: demographic-focused | Target Dataset: ma2019:

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



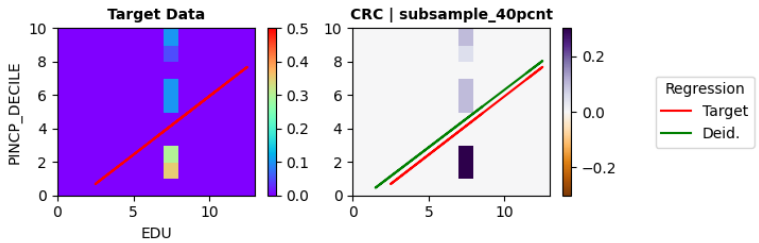
Feature Set: industry-focused | Target Dataset: ma2019:

Features: ['EDU', 'HISP', 'INDP_CAT', 'MSP', 'OWN_RENT', 'PINCP_DECILE', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 32,432,400



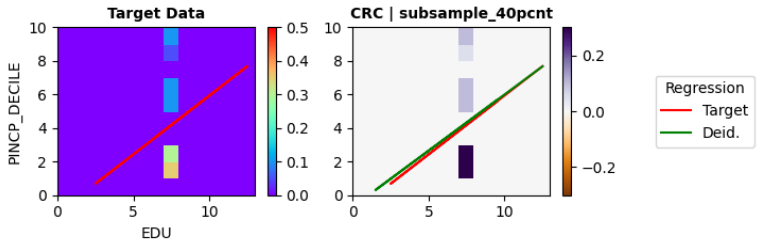
Feature Set: all-features | Target Dataset: ma2019:

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'PWGTP', 'RAC1P', 'SEX', 'WGTP']
Feature Space (possible combinations): 36,042,774,768,000,000,000,000,000



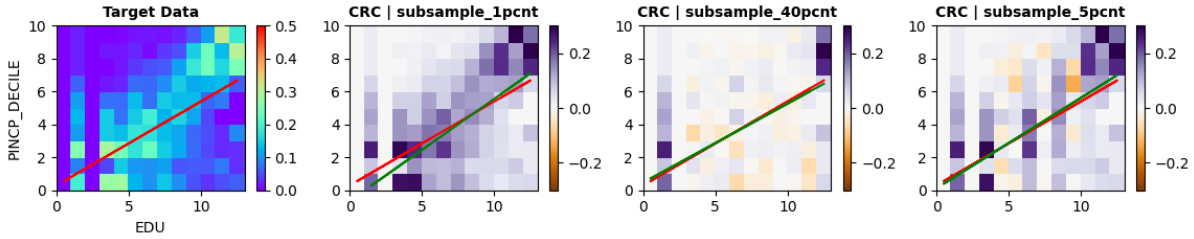
Feature Set: simple-features | Target Dataset: ma2019:

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 14,711,336,640,000,000,000



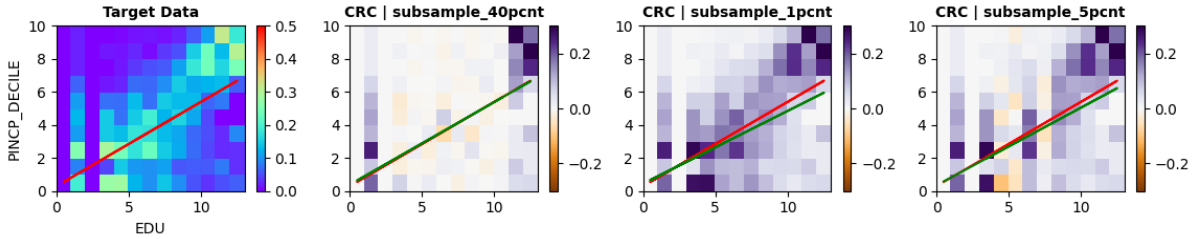
Feature Set: demographic-focused | Target Dataset: national2019:

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



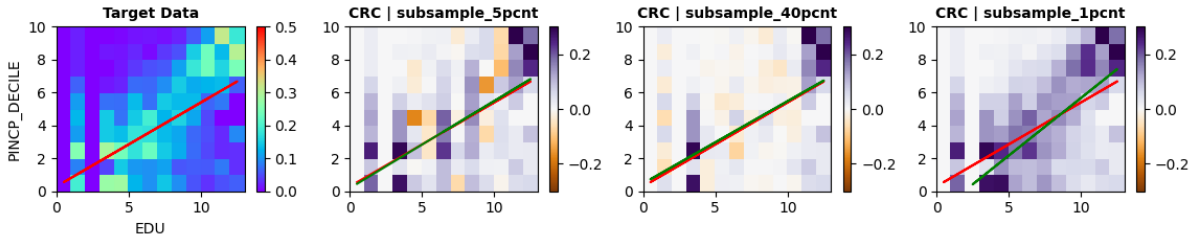
Feature Set: industry-focused | Target Dataset: national2019:

Features: ['EDU', 'HISP', 'INDP_CAT', 'MSP', 'OWN_RENT', 'PINCP_DECILE', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 32,432,400



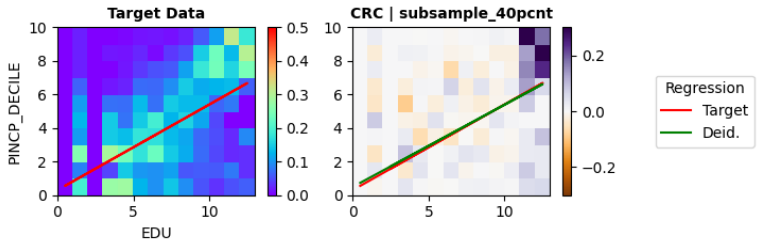
Feature Set: all-features | Target Dataset: national2019:

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'PWGTP', 'RAC1P', 'SEX', 'WGTP']
Feature Space (possible combinations): 36,042,774,768,000,000,000,000,000



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

Features: ['AGEP', 'DEAR', 'DENSITY', 'DEYE', 'DPHY', 'DREM', 'DVET', 'EDU', 'HISP', 'HOUSING_TYPE', 'INDP_CAT', 'MSP', 'NOC', 'NPF', 'OWN_RENT', 'PINCP', 'PINCP_DECILE', 'POVPIP', 'PUMA', 'RAC1P', 'SEX']
Feature Space (possible combinations): 14,711,336,640,000,000,000



Observations

How does subsampling impact utility? Find some interesting examples above and then for more detailed results on those variants, check out the detailed data reports zip file in the same folder as this report.

Are there other techniques in the archive that produce similar quality data to subsampling, but which provide better privacy?

Are there any techniques in the archive which provide better equity than subsampling?

Data Description

Deidentified (Deid.) Datasets:

| subsample_40pcnt | | :

Label Name	Label Value
Algorithm Name	subsample_40pcnt
Target Dataset	tx2019
Variant Label	
Algorithm Type	sdc
Library Name	subsample
Feature Set Name	demographic-focused
Privacy Category	sdc
Deid Data Id	f83e2f9b4d1dbcc865e6ee1af453cb3e4edc44f4
Features List	AGEP, SEX, MSP, RAC1P, HOUSING_TYPE, OWN_RENT, EDU, PINCP_DECILE, DVET, DEYE
Privacy Label Detail	A simple uniform random 40% subsample of the target data. It provides limited privacy by redacting records at random, allowing survey takers to deny that their information was included in the public release.
Submission Timestamp	5/20/2023 00:00:00
Team	CRC
Research Papers	https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf

Property	Value
Filename	subsample_40pcnt_demographic
Records	3710
Features	10

l subsample_40pcnt l l :

Label Name	Label Value
Algorithm Name	subsample_40pcnt
Target Dataset	ma2019
Variant Label	
Algorithm Type	sdc
Library Name	subsample
Feature Set Name	demographic-focused
Privacy Category	sdc
Deid Data Id	08b95430a7a380eb6325932a0a3ab6174a93e36c
Features List	AGEP, SEX, MSP, RAC1P, HOUSING_TYPE, OWN_RENT, EDU, PINCP_DECILE, DVET, DEYE
Privacy Label Detail	A simple uniform random 40% subsample of the target data. It provides limited privacy by redacting records at random, allowing survey takers to deny that their information was included in the public release.
Submission Timestamp	5/20/2023 00:00:00
Team	CRC
Research Papers	https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf

Property	Value
Filename	subsample_40pcnt_demographic
Records	3054
Features	10

l subsample_1pcnt l l :

Label Name	Label Value
Algorithm Name	subsample_1pcnt
Target Dataset	national2019
Variant Label	
Algorithm Type	sdc
Library Name	subsample
Feature Set Name	demographic-focused
Privacy Category	sdc
Privacy Label Detail	A simple uniform random 1% subsample of the target data. It provides limited privacy by redacting records at random, allowing survey takers to deny that their information was included in the public release.
Features List	AGEP, SEX, MSP, RAC1P, HOUSING_TYPE, OWN_RENT, EDU, PINCP_DECILE, DVET, DEYE
Deid Data Id	e4ff5a9738ee73cb740ec7799f5bd8c6ec790425
Submission Timestamp	6/16/2023 00:00:00
Team	CRC
Research Papers	https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf

Property	Value
Filename	subsample_1pcnt_demographic
Records	273
Features	10

| subsample_40pcnt | | :

Label Name	Label Value
Algorithm Name	subsample_40pcnt
Target Dataset	national2019
Variant Label	
Algorithm Type	sdc
Library Name	subsample
Feature Set Name	demographic-focused
Privacy Category	sdc
Deid Data Id	91293c6fb1be4e7d3aa6340b0658317218f70f2e
Features List	AGEP, SEX, MSP, RAC1P, HOUSING_TYPE, OWN_RENT, EDU, PINCP_DECILE, DVET, DEYE
Privacy Label Detail	A simple uniform random 40% subsample of the target data. It provides limited privacy by redacting records at random, allowing survey takers to deny that their information was included in the public release.
Submission Timestamp	5/20/2023 00:00:00
Team	CRC
Research Papers	https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf

Property	Value
Filename	subsample_40pcnt_demographic
Records	10901
Features	10

| subsample_5pcnt | | :

Label Name	Label Value
Algorithm Name	subsample_5pcnt
Target Dataset	national2019
Variant Label	
Algorithm Type	sdc
Library Name	subsample
Feature Set Name	demographic-focused
Privacy Category	sdc
Privacy Label Detail	A simple uniform random 5% subsample of the target data. It provides limited privacy by redacting records at random, allowing survey takers to deny that their information was included in the public release.
Features List	AGEP, SEX, MSP, RAC1P, HOUSING_TYPE, OWN_RENT, EDU, PINCP_DECILE, DVET, DEYE
Deid Data Id	8be2b4bd5c86a0206b3eee3fb2168fbcf5018d17
Submission Timestamp	6/16/2023 00:00:00
Team	CRC
Research Papers	https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf

Property	Value
Filename	subsample_5pcnt_demographic
Records	1363
Features	10

l subsample_40pcnt l l :

Label Name	Label Value
Algorithm Name	subsample_40pcnt
Target Dataset	national2019
Variant Label	
Algorithm Type	sdc
Library Name	subsample
Feature Set Name	industry-focused
Privacy Category	sdc
Deid Data Id	708868a8cec27dfbc0469707d8b7309c84608803
Features List	PUMA, SEX, MSP, HISP, RAC1P, OWN_RENT, INDP_CAT, EDU, PINCP_DECILE
Privacy Label Detail	A simple uniform random 40% subsample of the target data. It provides limited privacy by redacting records at random, allowing survey takers to deny that their information was included in the public release.
Submission Timestamp	5/20/2023 00:00:00
Team	CRC
Research Papers	https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf

Property	Value
Filename	subsample_40pcnt_industry
Records	10901
Features	9

l subsample_1pcnt l l :

Label Name	Label Value
Algorithm Name	subsample_1pcnt
Target Dataset	national2019
Variant Label	
Algorithm Type	sdc
Library Name	subsample
Feature Set Name	industry-focused
Privacy Category	sdc
Privacy Label Detail	A simple uniform random 1% subsample of the target data. It provides limited privacy by redacting records at random, allowing survey takers to deny that their information was included in the public release.
Features List	PUMA, SEX, MSP, HISP, RAC1P, OWN_RENT, INDP_CAT, EDU, PINCP_DECILE
Deid Data Id	a51cfd4103b0ee81691de6d9bb1e83f03956ec3c
Submission Timestamp	6/16/2023 00:00:00
Team	CRC
Research Papers	https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf

Property	Value
Filename	subsample_1pcnt_industry
Records	273
Features	9

l subsample_40pcnt l l :

Label Name	Label Value
Algorithm Name	subsample_40pcnt
Target Dataset	ma2019
Variant Label	
Algorithm Type	sdc
Library Name	subsample
Feature Set Name	industry-focused
Privacy Category	sdc
Deid Data Id	107ed066e79172f590c021631c792b7a94727f01
Features List	PUMA, SEX, MSP, HISP, RAC1P, OWN_RENT, INDP_CAT, EDU, PINCP_DECILE
Privacy Label Detail	A simple uniform random 40% subsample of the target data. It provides limited privacy by redacting records at random, allowing survey takers to deny that their information was included in the public release.
Submission Timestamp	5/20/2023 00:00:00
Team	CRC
Research Papers	https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf

Property	Value
Filename	subsample_40pcnt_industry
Records	3054
Features	9

l subsample_5pcnt l l :

Label Name	Label Value
Algorithm Name	subsample_5pcnt
Target Dataset	national2019
Variant Label	
Algorithm Type	sdc
Library Name	subsample
Feature Set Name	industry-focused
Privacy Category	sdc
Privacy Label Detail	A simple uniform random 5% subsample of the target data. It provides limited privacy by redacting records at random, allowing survey takers to deny that their information was included in the public release.
Features List	PUMA, SEX, MSP, HISP, RAC1P, OWN_RENT, INDP_CAT, EDU, PINCP_DECILE
Deid Data Id	e74a024ec17acb19a1877a22c699b36cbb08b052
Submission Timestamp	6/16/2023 00:00:00
Team	CRC
Research Papers	https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf

Property	Value
Filename	subsample_5pcnt_industry
Records	1363
Features	9

| subsample_40pcnt | | :

Label Name	Label Value
Algorithm Name	subsample_40pcnt
Target Dataset	tx2019
Variant Label	
Algorithm Type	sdc
Library Name	subsample
Feature Set Name	industry-focused
Privacy Category	sdc
Deid Data Id	b7112651b7bb812d9fe5bb3db2320012c213f5e9
Features List	PUMA, SEX, MSP, HISP, RAC1P, OWN_RENT, INDP_CAT, EDU, PINCP_DECILE
Privacy Label Detail	A simple uniform random 40% subsample of the target data. It provides limited privacy by redacting records at random, allowing survey takers to deny that their information was included in the public release.
Submission Timestamp	5/20/2023 00:00:00
Team	CRC
Research Papers	https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf

Property	Value
Filename	subsample_40pcnt_industry
Records	3710
Features	9

| subsample_40pcnt | | :

Label Name	Label Value
Algorithm Name	subsample_40pcnt
Target Dataset	ma2019
Variant Label	
Algorithm Type	sdc
Library Name	subsample
Feature Set Name	family-focused
Privacy Category	sdc
Deid Data Id	28bd61ace25aeb68591c02f6fd07abee74796e26
Features List	PUMA, AGEP, SEX, MSP, HISP, RAC1P, NOC, NPF, OWN_RENT, PINCP_DECILE, POVPIP
Privacy Label Detail	A simple uniform random 40% subsample of the target data. It provides limited privacy by redacting records at random, allowing survey takers to deny that their information was included in the public release.
Submission Timestamp	5/20/2023 00:00:00
Team	CRC
Research Papers	https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf

Property	Value
Filename	subsample_40pcnt_family
Records	3054
Features	11

| subsample_1pcnt | | :

Label Name	Label Value
Algorithm Name	subsample_1pcnt
Target Dataset	national2019
Variant Label	
Algorithm Type	sdc
Library Name	subsample
Feature Set Name	family-focused
Privacy Category	sdc
Privacy Label Detail	A simple uniform random 1% subsample of the target data. It provides limited privacy by redacting records at random, allowing survey takers to deny that their information was included in the public release.
Features List	PUMA, AGEP, SEX, MSP, HISP, RAC1P, NOC, NPF, OWN_RENT, PINCP_DECILE, POVPIP
Deid Data Id	329febcfd479c788ee1aa0d6b8b7f0f89d60b021
Submission Timestamp	6/16/2023 00:00:00
Team	CRC
Research Papers	https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf

Property	Value
Filename	subsample_1pcnt_family
Records	273
Features	11

| subsample_40pcnt | | :

Label Name	Label Value
Algorithm Name	subsample_40pcnt
Target Dataset	national2019
Variant Label	
Algorithm Type	sdc
Library Name	subsample
Feature Set Name	family-focused
Privacy Category	sdc
Deid Data Id	fe89370246500657e2463ef98db8a992717327cc
Features List	PUMA, AGEP, SEX, MSP, HISP, RAC1P, NOC, NPF, OWN_RENT, PINCP_DECILE, POVPIP
Privacy Label Detail	A simple uniform random 40% subsample of the target data. It provides limited privacy by redacting records at random, allowing survey takers to deny that their information was included in the public release.
Submission Timestamp	5/20/2023 00:00:00
Team	CRC
Research Papers	https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf

Property	Value
Filename	subsample_40pcnt_family
Records	10901
Features	11

| subsample_5pcnt | | :

Label Name	Label Value
Algorithm Name	subsample_5pcnt
Target Dataset	national2019
Variant Label	
Algorithm Type	sdc
Library Name	subsample
Feature Set Name	family-focused
Privacy Category	sdc
Privacy Label Detail	A simple uniform random 5% subsample of the target data. It provides limited privacy by redacting records at random, allowing survey takers to deny that their information was included in the public release.
Features List	PUMA, AGEP, SEX, MSP, HISP, RAC1P, NOC, NPF, OWN_RENT, PINCP_DECILE, POVPIP
Deid Data Id	c88c09340dcadb6c8258627c503d33cefca6c679
Submission Timestamp	6/16/2023 00:00:00
Team	CRC
Research Papers	https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf

Property	Value
Filename	subsample_5pcnt_family
Records	1363
Features	11

| subsample_40pcnt | | :

Label Name	Label Value
Algorithm Name	subsample_40pcnt
Target Dataset	tx2019
Variant Label	
Algorithm Type	sdc
Library Name	subsample
Feature Set Name	family-focused
Privacy Category	sdc
Deid Data Id	5fd422b23b3b8f022af9b05d7d531decda11b64a
Features List	PUMA, AGEP, SEX, MSP, HISP, RAC1P, NOC, NPF, OWN_RENT, PINCP_DECILE, POVPIP
Privacy Label Detail	A simple uniform random 40% subsample of the target data. It provides limited privacy by redacting records at random, allowing survey takers to deny that their information was included in the public release.
Submission Timestamp	5/20/2023 00:00:00
Team	CRC
Research Papers	https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf

Property	Value
Filename	subsample_40pcnt_family
Records	3710
Features	11

| subsample_5pcnt | | :

Label Name	Label Value
Algorithm Name	subsample_5pcnt
Target Dataset	national2019
Variant Label	
Algorithm Type	sdc
Library Name	subsample
Feature Set Name	all-features
Privacy Category	sdc
Privacy Label Detail	A simple uniform random 5% subsample of the target data. It provides limited privacy by redacting records at random, allowing survey takers to deny that their information was included in the public release.
Features List	PUMA, AGEP, SEX, MSP, HISP, RAC1P, NOC, NPF, HOUSING_TYPE, OWN_RENT, DENSITY, INDP, INDP_CAT, EDU, PINCP, PINCP_DECILE, POVPIP, DVET, DREM, DPHY, DEYE, DEAR, WGTP, PWGTP
Deid Data Id	65a1e373caf4e80379d40560924e5756b57cda23
Submission Timestamp	6/16/2023 00:00:00
Team	CRC
Research Papers	https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf

Property	Value
Filename	subsample_5pcnt_all
Records	1363
Features	24

| subsample_40pcnt | | :

Label Name	Label Value
Algorithm Name	subsample_40pcnt
Target Dataset	national2019
Variant Label	
Algorithm Type	sdc
Library Name	subsample
Feature Set Name	all-features
Privacy Category	sdc
Deid Data Id	434816825fd1677cd20097cbcb59bbe6870a7b0d
Features List	PUMA, AGEP, SEX, MSP, HISP, RAC1P, NOC, NPF, HOUSING_TYPE, OWN_RENT, DENSITY, INDP, INDP_CAT, EDU, PINCP, PINCP_DECILE, POVPIP, DVET, DREM, DPHY, DEYE, DEAR, WGTP, PWGTP
Privacy Label Detail	A simple uniform random 40% subsample of the target data. It provides limited privacy by redacting records at random, allowing survey takers to deny that their information was included in the public release.
Submission Timestamp	5/20/2023 00:00:00
Team	CRC
Research Papers	https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf

Property	Value
Filename	subsample_40pcnt_all
Records	10901
Features	24

| subsample_1pcnt | | :

Label Name	Label Value
Algorithm Name	subsample_1pcnt
Target Dataset	national2019
Variant Label	
Algorithm Type	sdc
Library Name	subsample
Feature Set Name	all-features
Privacy Category	sdc
Privacy Label Detail	A simple uniform random 1% subsample of the target data. It provides limited privacy by redacting records at random, allowing survey takers to deny that their information was included in the public release.
Features List	PUMA, AGEP, SEX, MSP, HISP, RAC1P, NOC, NPF, HOUSING_TYPE, OWN_RENT, DENSITY, INDP, INDP_CAT, EDU, PINCP, PINCP_DECILE, POVPIP, DVET, DREM, DPHY, DEYE, DEAR, WGTP, PWGTP
Deid Data Id	7f6fd7c92c0e4c1d891eaea30d5cc9291da06c0e
Submission Timestamp	6/16/2023 00:00:00
Team	CRC
Research Papers	https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf

Property	Value
Filename	subsample_1pcnt_all
Records	273
Features	24

| subsample_40pcnt | | :

Label Name	Label Value
Algorithm Name	subsample_40pcnt
Target Dataset	tx2019
Variant Label	
Algorithm Type	sdc
Library Name	subsample
Feature Set Name	all-features
Privacy Category	sdc
Deid Data Id	1f0933fa2d6778126b2a986199e866afb5f9a035
Features List	PUMA, AGEP, SEX, MSP, HISP, RAC1P, NOC, NPF, HOUSING_TYPE, OWN_RENT, DENSITY, INDP, INDP_CAT, EDU, PINCP, PINCP_DECILE, POVPIP, DVET, DREM, DPHY, DEYE, DEAR, WGTP, PWGTP
Privacy Label Detail	A simple uniform random 40% subsample of the target data. It provides limited privacy by redacting records at random, allowing survey takers to deny that their information was included in the public release.
Submission Timestamp	5/20/2023 00:00:00
Team	CRC
Research Papers	https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf

Property	Value
Filename	subsample_40pcnt_all
Records	3710
Features	24

l subsample_40pcnt l l :

Label Name	Label Value
Algorithm Name	subsample_40pcnt
Target Dataset	ma2019
Variant Label	
Algorithm Type	sdc
Library Name	subsample
Feature Set Name	all-features
Privacy Category	sdc
Deid Data Id	f12e8ac0afef10b14f09bd9e41af2b5626e0f6f6
Features List	PUMA, AGEP, SEX, MSP, HISP, RAC1P, NOC, NPF, HOUSING_TYPE, OWN_RENT, DENSITY, INDP, INDP_CAT, EDU, PINCP, PINCP_DECILE, POVPIP, DVET, DREM, DPHY, DEYE, DEAR, WGTP, PWGTP
Privacy Label Detail	A simple uniform random 40% subsample of the target data. It provides limited privacy by redacting records at random, allowing survey takers to deny that their information was included in the public release.
Submission Timestamp	5/20/2023 00:00:00
Team	CRC
Research Papers	https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf

Property	Value
Filename	subsample_40pcnt_all
Records	3054
Features	24

l subsample_40pcnt l l :

Label Name	Label Value
Algorithm Name	subsample_40pcnt
Target Dataset	ma2019
Variant Label	
Algorithm Type	sdc
Library Name	subsample
Feature Set Name	simple-features
Privacy Category	sdc
Deid Data Id	b57f74ff366799e215c650a5ab5c1fbd939bf36d
Features List	PUMA, AGEP, SEX, MSP, HISP, RAC1P, NOC, NPF, HOUSING_TYPE, OWN_RENT, DENSITY, INDP_CAT, EDU, PINCP, PINCP_DECILE, POVPIP, DVET, DREM, DPHY, DEYE, DEAR
Privacy Label Detail	A simple uniform random 40% subsample of the target data. It provides limited privacy by redacting records at random, allowing survey takers to deny that their information was included in the public release.
Submission Timestamp	5/20/2023 00:00:00
Team	CRC
Research Papers	https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf

Property	Value
Filename	subsample_40pcnt_simple
Records	3054
Features	21

| subsample_40pcnt | | :

Label Name	Label Value
Algorithm Name	subsample_40pcnt
Target Dataset	national2019
Variant Label	
Algorithm Type	sdc
Library Name	subsample
Feature Set Name	simple-features
Privacy Category	sdc
Deid Data Id	cfb5d8d6f93b1d7528371e52073a4566716291cb
Features List	PUMA, AGEP, SEX, MSP, HISP, RAC1P, NOC, NPF, HOUSING_TYPE, OWN_RENT, DENSITY, INDP_CAT, EDU, PINCP, PINCP_DECILE, POVPIP, DVET, DREM, DPHY, DEYE, DEAR
Privacy Label Detail	A simple uniform random 40% subsample of the target data. It provides limited privacy by redacting records at random, allowing survey takers to deny that their information was included in the public release.
Submission Timestamp	5/20/2023 00:00:00
Team	CRC
Research Papers	https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf

Property	Value
Filename	subsample_40pcnt_simple
Records	10901
Features	21

| subsample_40pcnt | | :

Label Name	Label Value
Algorithm Name	subsample_40pcnt
Target Dataset	tx2019
Variant Label	
Algorithm Type	sdc
Library Name	subsample
Feature Set Name	simple-features
Privacy Category	sdc
Deid Data Id	77c515fcc66748051f1ccbd6cf8cd4838bada8ce
Features List	PUMA, AGEP, SEX, MSP, HISP, RAC1P, NOC, NPF, HOUSING_TYPE, OWN_RENT, DENSITY, INDP_CAT, EDU, PINCP, PINCP_DECILE, POVPIP, DVET, DREM, DPHY, DEYE, DEAR
Privacy Label Detail	A simple uniform random 40% subsample of the target data. It provides limited privacy by redacting records at random, allowing survey takers to deny that their information was included in the public release.
Submission Timestamp	5/20/2023 00:00:00
Team	CRC
Research Papers	https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021_ch01.pdf

Property	Value
Filename	subsample_40pcnt_simple
Records	3710
Features	21

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 | 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 | 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 | 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 | 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 | 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 | 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 | Bin Range: (6243.68, 9589.66]

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

Density Bin: 10 | 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 | 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 | 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 | 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