YData

SYNTHETIC DATA QUALITY REPORT

10K # ORIGINAL EVENTS

10K # NEW EVENTS GENERATED

0 # MISSING VALUES DETECTED

0.81

UTILITY SCORE

2.3%

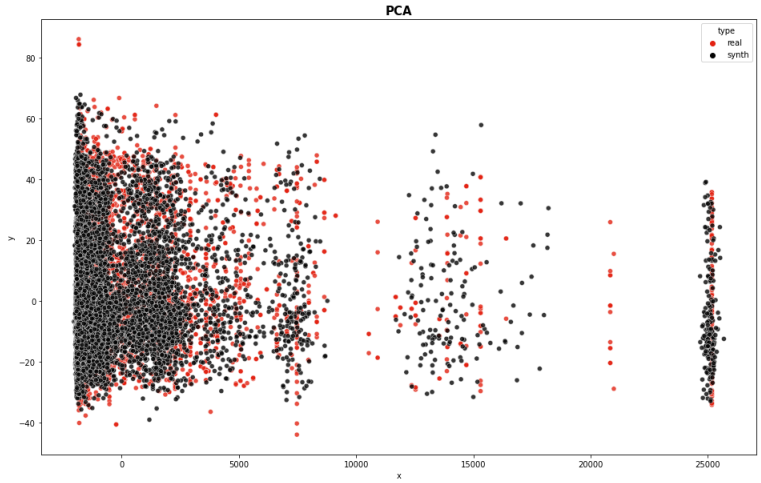
PRIVACY SCORE

1.32

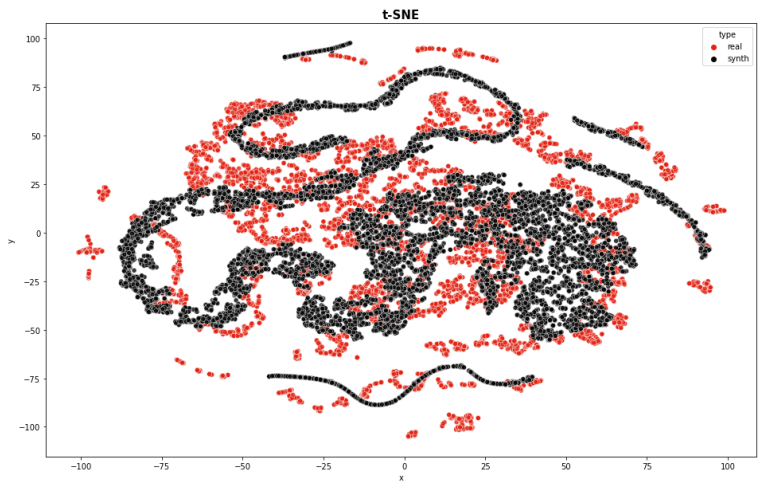
FIDELITY SCORE

DIMENSIONALITY REDUCTION

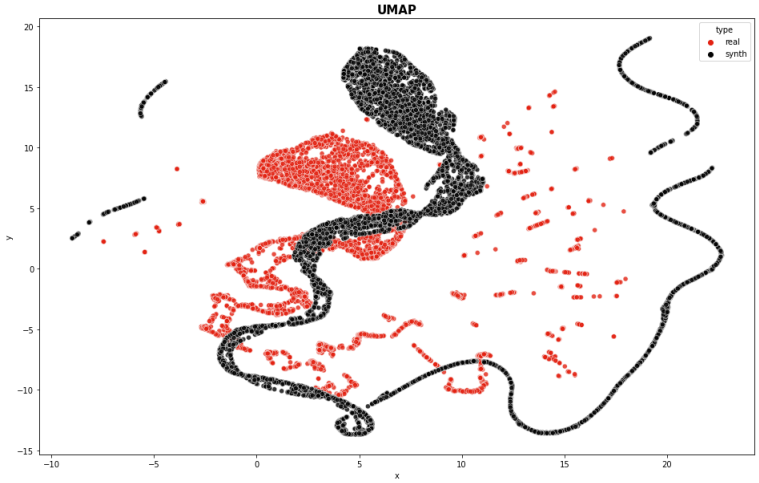
The t-SNE, PCA and UMAP dimensionality reduction techniques allow us to visually compare distributions of synthetic and real data on humanly perceptible dimensions (e.g. 2D). Although these techniques are based on different principles, the underlying assumption is that if the distributions do not differ heavily with regard to their origin (synthetic vs real), the utility of synthetic data will be large. Any fundamental difference in the distributions of data would be captured by the dimensionality reduction techniques and such difference would be reflected by different visual structure of scatterplots cluster points.



PCA: principal component analysis



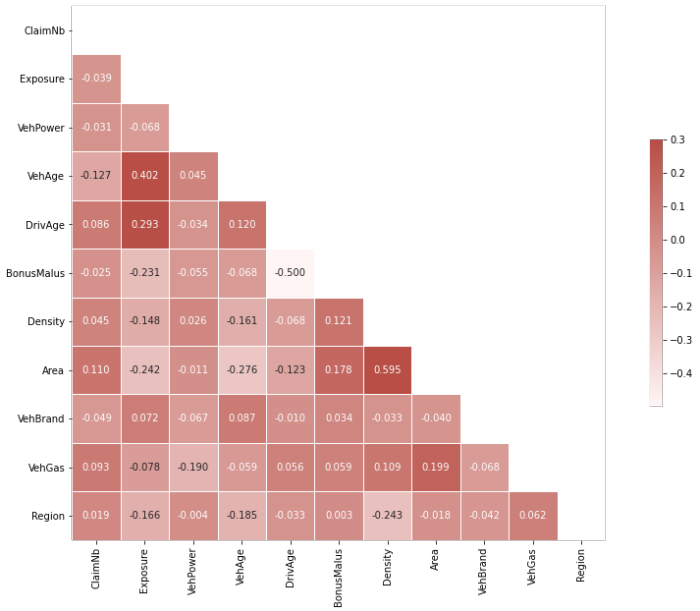
t-SNE: t-distributed stochastic neighbor embedding



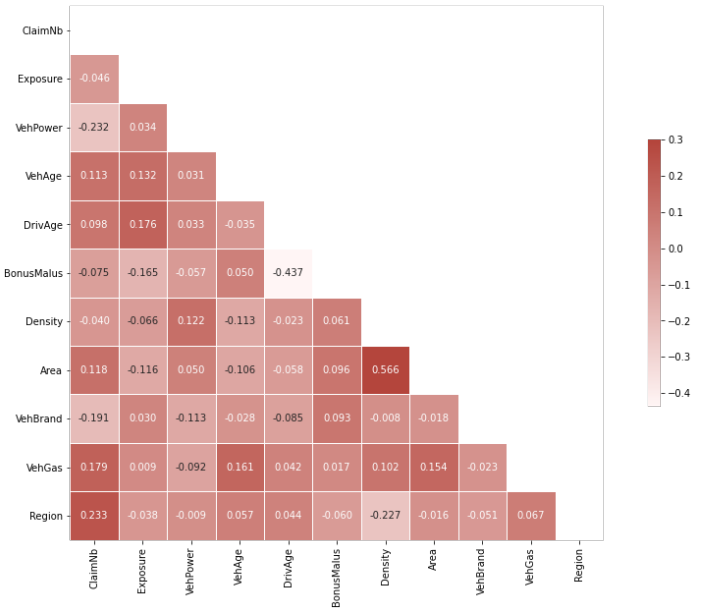
UMAP: a non-linear dimensionality reduction technique

CORRELATION HEATMAPS

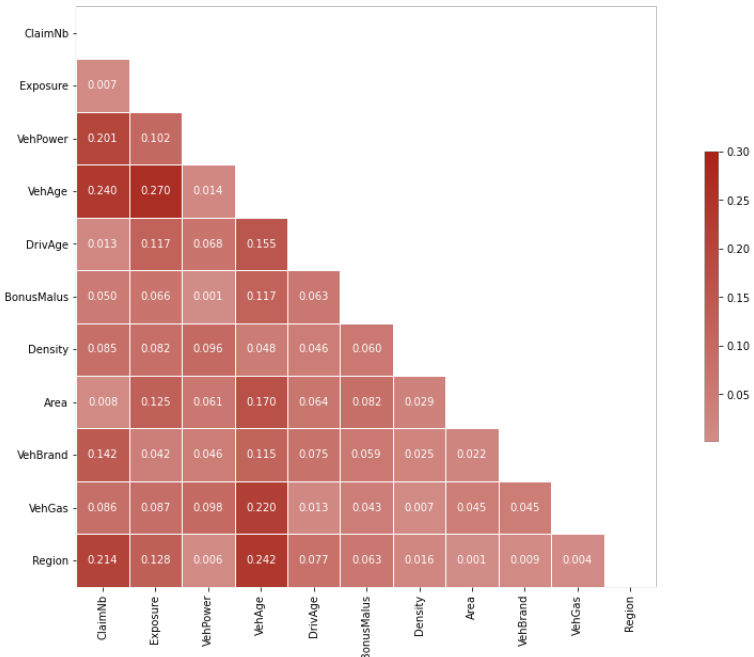
The correlation heatmaps provide a measure of association between features. On the top row, the correlations provide the association between the features within the same datasets (real on real, synthetic on synthetic). On the bottom row, the difference between real and synthetic correlation matrices provides an indication to where the synthesization process is deviating more from the original data. Larger absolute differences mean smaller utility.



Correlation between features in real data.



Correlation between features in synthetic data.



Differences between correlations in real and synthetic data.

FIDELITY METRICS

Fidelity measures quantify the consistency between results achieved with synthetic and real data. Performance is measured by '('mean absolute error',)' for a 'regression' task with the target on the 'Exposure' feature.

0.49

TSTR

0.37

TRTR

Estimators	Real Data	Synth Data
Linear Regression	0.3	0.3
Multi-layer Perceptron	0.8	1.5
Decision Tree	0.3	0.4
Ridge	0.3	0.3
Lasso	0.3	0.3
Linear Support Vector	0.3	0.3

Real Data performances use a **TRTR** (Train Real, Test Real) approach, providing the original performance of real data. Synth Data performances use a **TSTR** (Train Synthetic, Test Real) approach, providing the performance of synthetic data on real targets.
While TRTR measures the feasibility of the underlying use case, TSTR provides the feasibility of substituting the original data with synthetically generated samples.

UTILITY METRICS

Utility measures quantify the consistency of inherent properties between the real and synthetic data.

1.1	764	2M	0.81
DISTANCE CORRELATION	DISTANCE STATISTICS	DISTANCE COVARIANCE	DISTANCE DISTRIBUTION

The **DISTANCE CORRELATION** metric measures the correlation (i.e. strengh of association) between features of real and synthetic data. Larger values mean higher utility.
The **DISTANCE STATISTICS** metric measures the euclidean distance between statistical properties (e.g. mean, standard deviation, min, max, quantiles) of synthetic and real data. Smaller values mean higher utility.
The **DISTANCE COVARIANCE** metric measures the euclidean distance between synthetic and real datasets' feature covariances matrices. Smaller values mean higher utility.
The **DISTANCE DISTRIBUTION** metric measures the similarity between the distributions of the features between real and synthetic data. Larger values mean higher utility.

PRIVACY METRICS

Privacy measures quantify the degree of exposure of real samples carried within the synthetically generated data.

0	0.98	14.49	0.84	0.0
EXACT MATCHES	SYNTH CLASSIFIER	PRIVACY AT RISK	HAMMING PRIVACY	NEIGHBOURS PRIVACY

The **EXACT MATCHES** provide the number of equal records between synthetic and real data. Smaller values mean higher privacy.
The **SYNTH CLASSIFIER** metric provides the performance (in ROC-AUC) of a model trained to distinguish real from synthetic data. Smaller values means higher privacy.
The **PRIVACY AT RISK** metric provides the percentage of the records that can potentially be re-identified back to the original information. Smaller values means higher privacy.
The **HAMMING PRIVACY** measures the average Hamming distance between synthetic and real data. Larger values means higher privacy.
The **NEIGHBOURS PRIVACY** provides the percentage of Nearest Neighbors in synthetic data which are too close to real data. Smaller values means higher privacy.