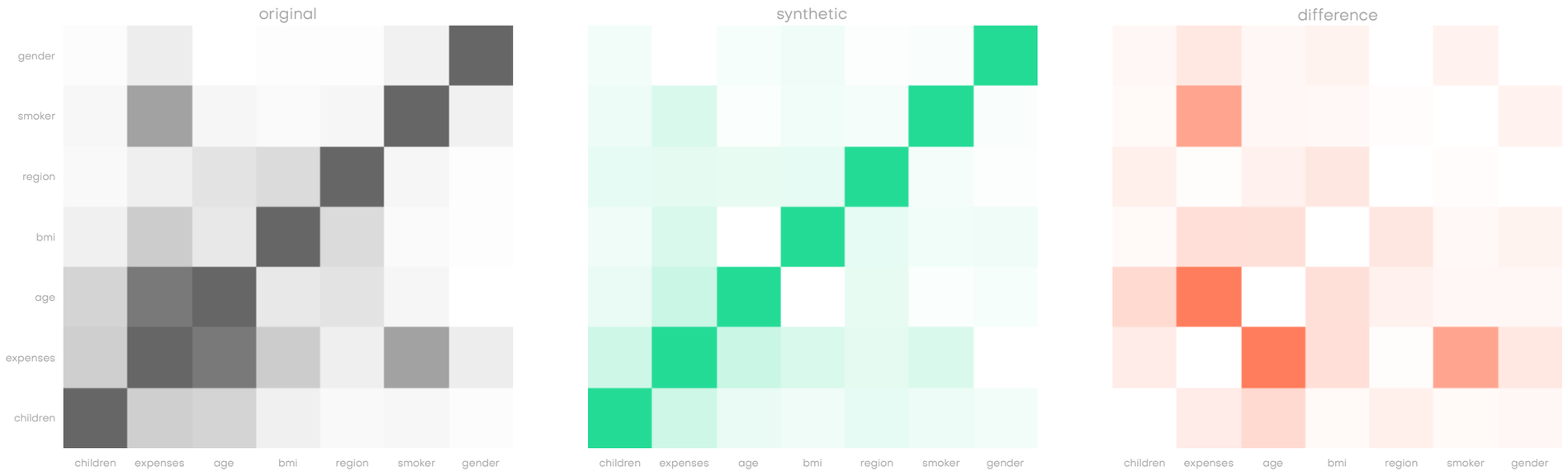
Model Report for `insurance_dataset`

Generated on 24 Jul 2024, 23:00 • Generator: e9803b10-c368-4d5b-958f-22dfa1c13ce3

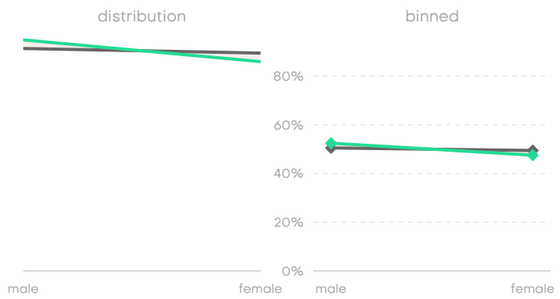
Dataset		Accuracy		Distances	
Original Samples	1,338	92.0% (94.6%)	Univariate 95.8%	Identical Matches	0.0% (0.1%)
Synthetic Samples	1,338		Bivariate 88.2%	Average Distances	2.19 (1.28)
Target Columns	7				

Correlations

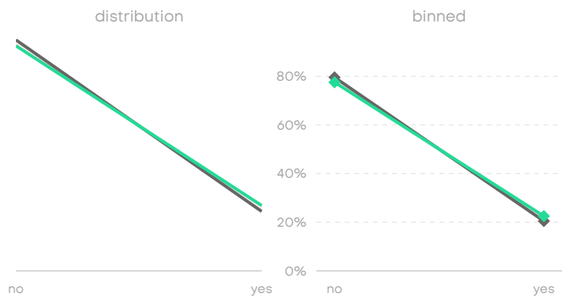
Correlation Matrices



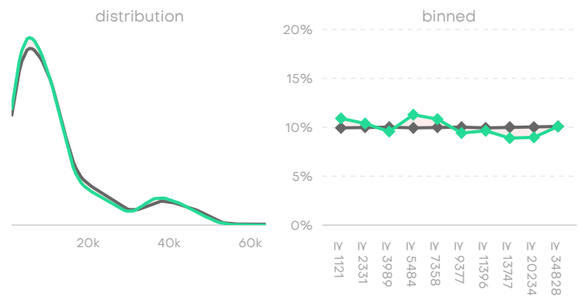
gender 98.1%



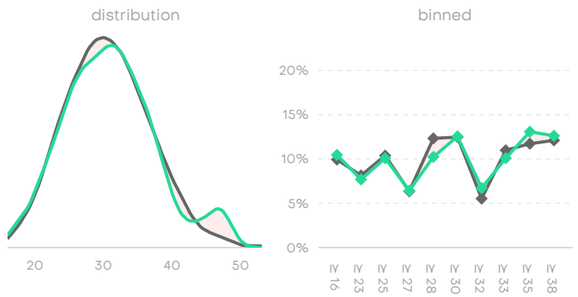
smoker 98.0%



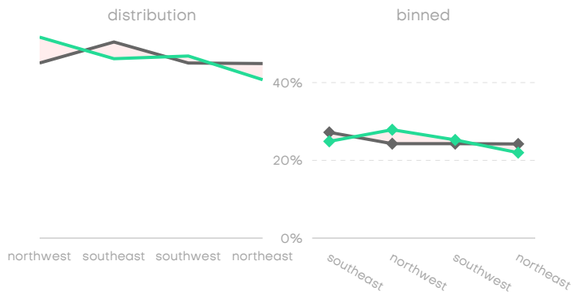
expenses 96.5%



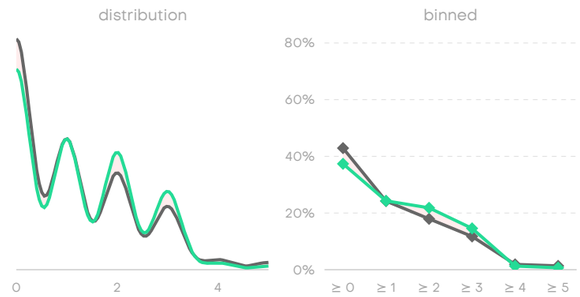
bmi 96.3%



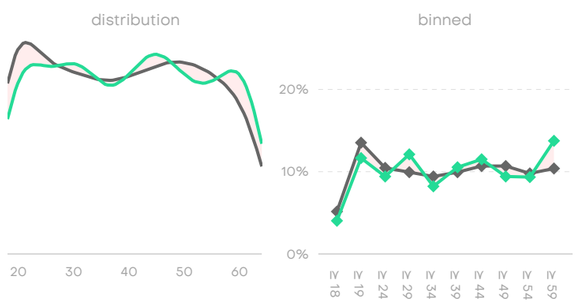
region 95.4%



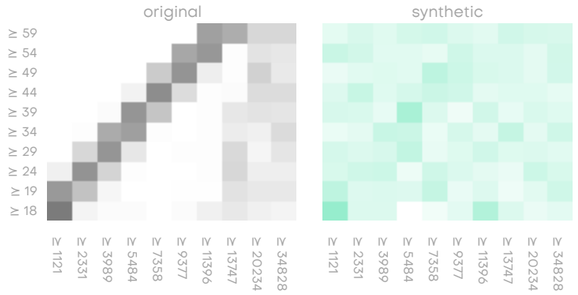
children 93.2%

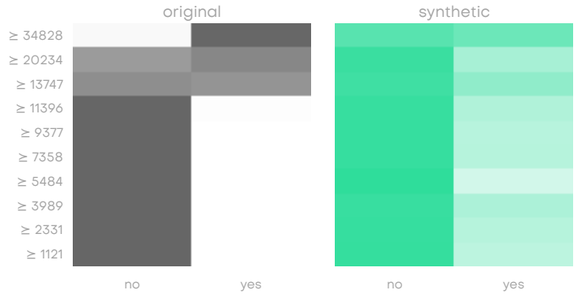


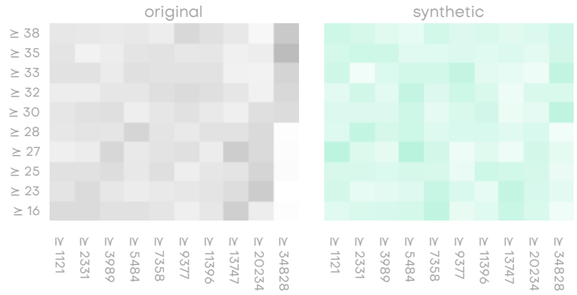
age 93.0%



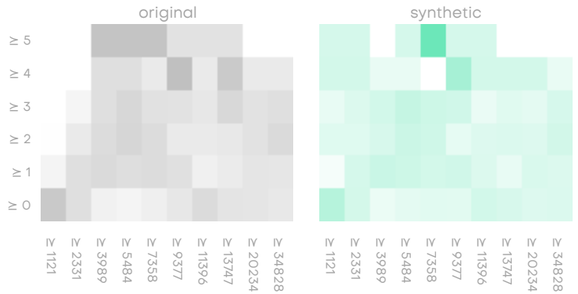
Bivariate Distributions



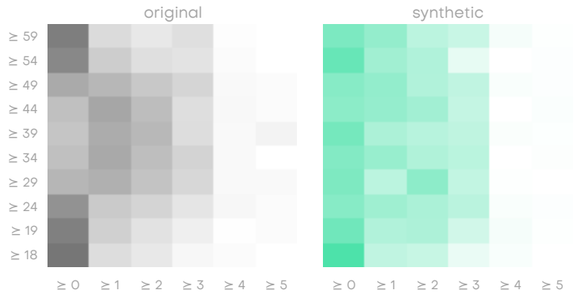




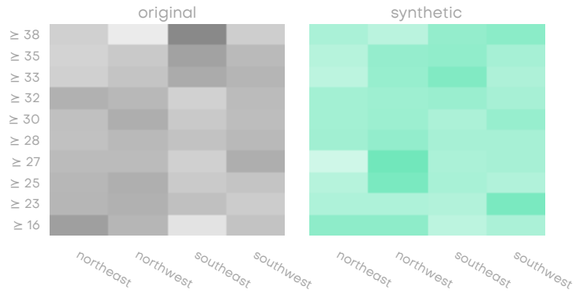
children ~ expenses 89.2%



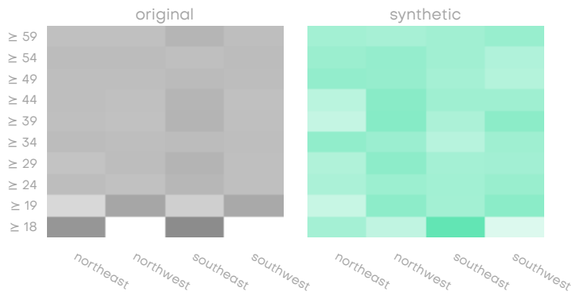
age ~ children 82.1%



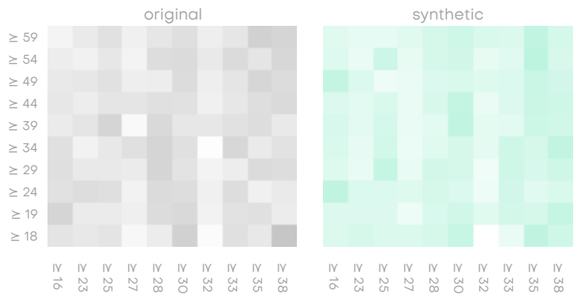
bmi ~ region 87.6%



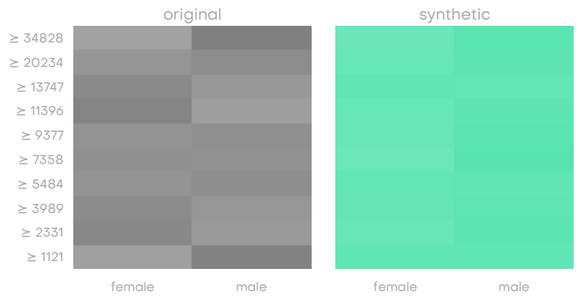
age ~ region 89.5%



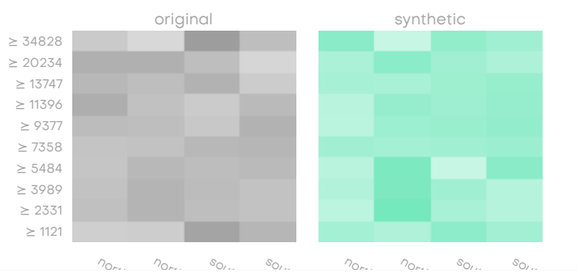
age ~ bmi 84.2%



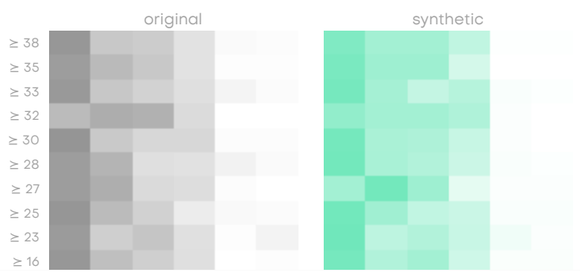
expenses ~ gender 92.5%



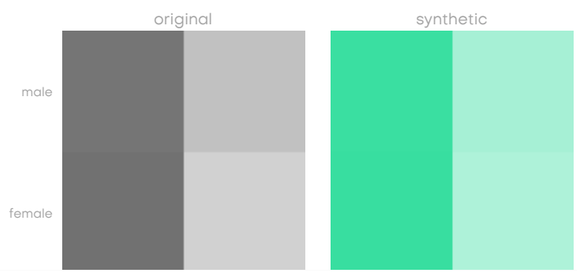
expenses ~ region 90.4%



bmi ~ children 89.2%

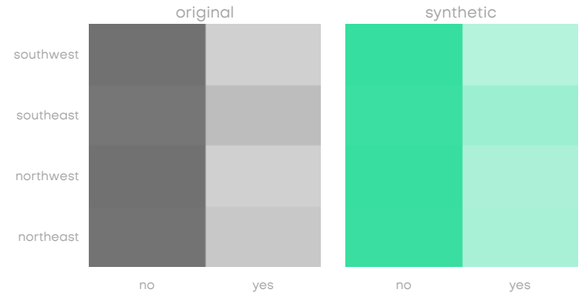


gender ~ smoker 96.5%

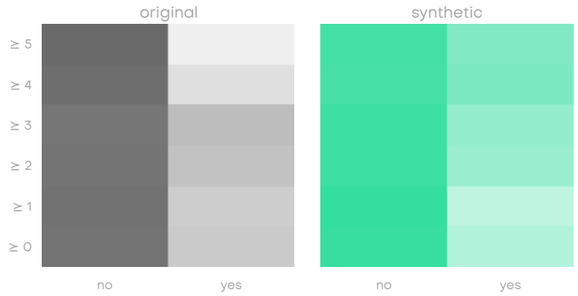


age ~ smoker 91.3%

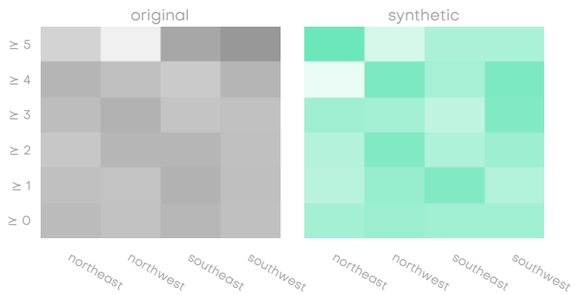
region ~ smoker 95.4%



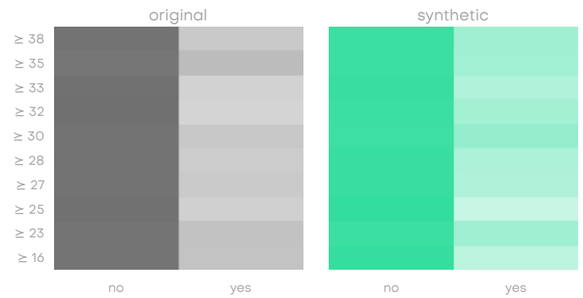
children ~ smoker 92.5%



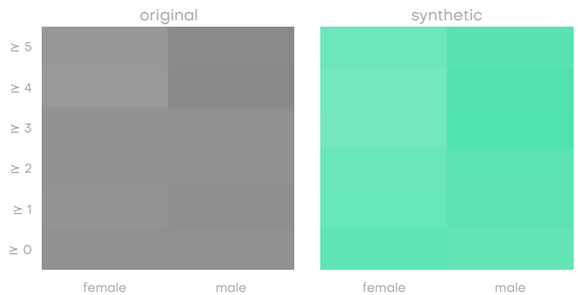
children ~ region 90.5%



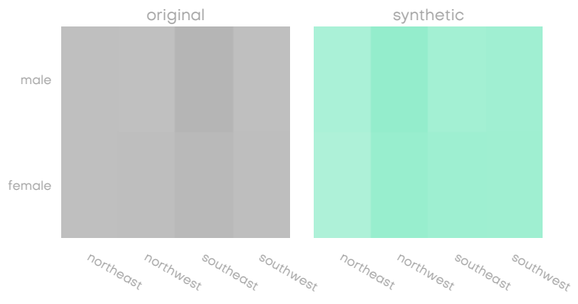
bmi ~ smoker 94.5%



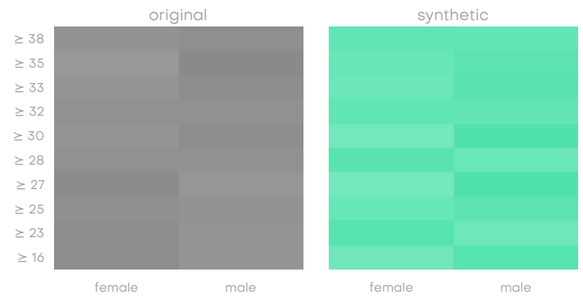
children ~ gender 92.8%



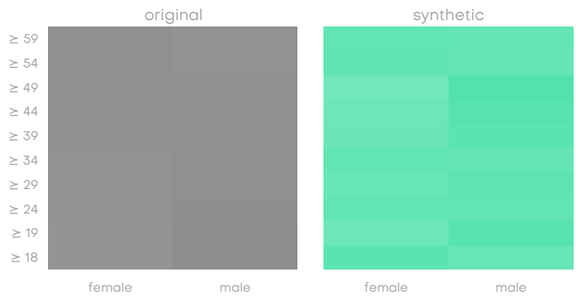
gender ~ region 95.4%



bmi ~ gender 93.6%

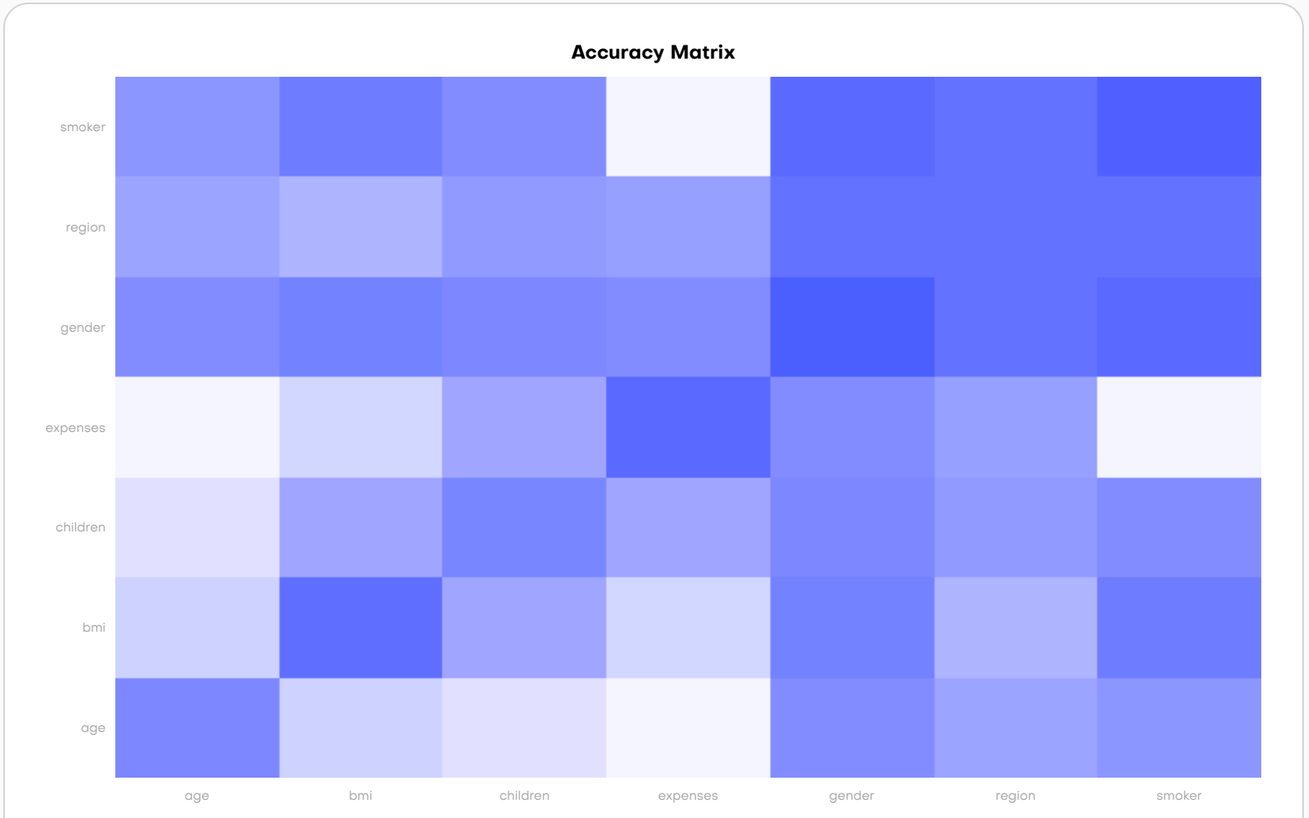


age ~ gender 92.4%



Accuracy

smoker	98.0%	91.3%
expenses	96.5%	80.7%
bmi	96.3%	88.8%
region	95.4%	91.5%
children	93.2%	89.4%
age	93.0%	81.7%
Total	**95.8%**	**88.2%**

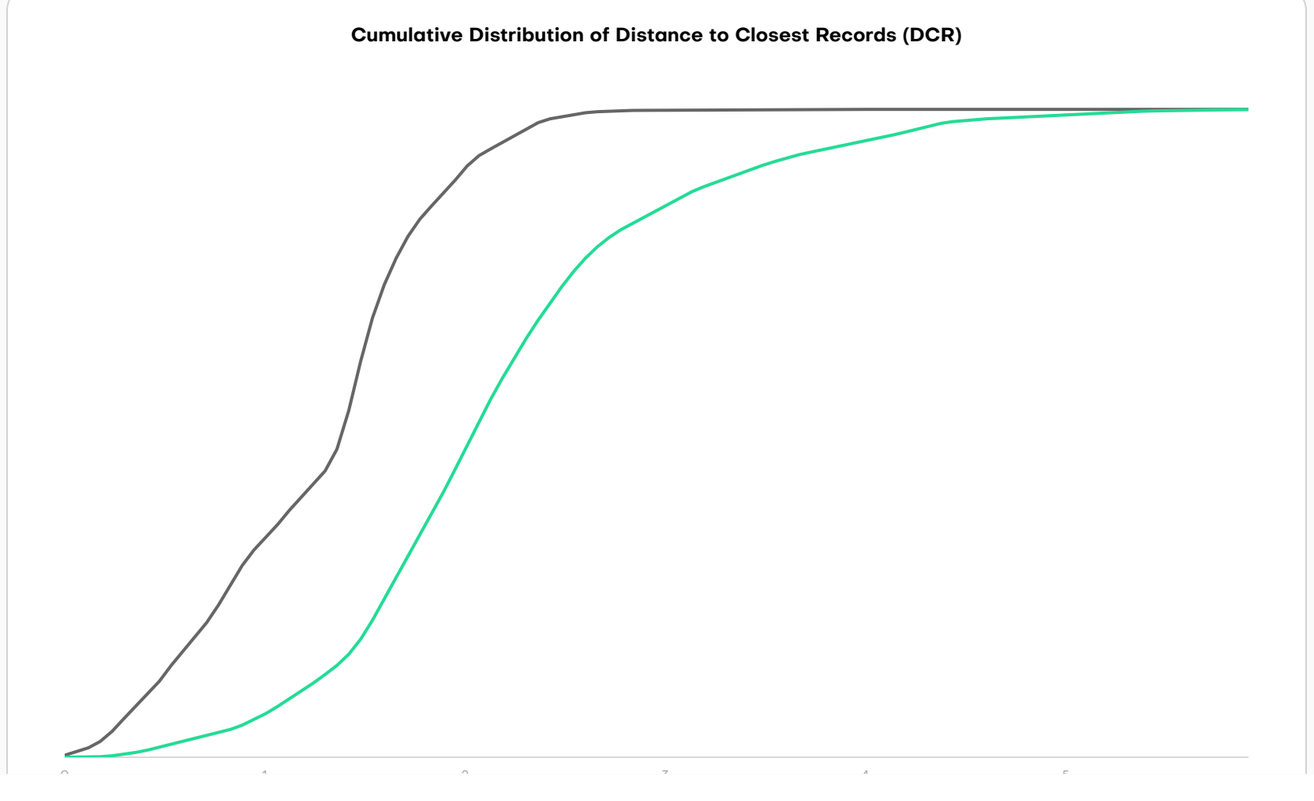


Accuracy of synthetic data is assessed by comparing the distributions of the synthetic (shown in green) and the original data (shown in gray). For each distribution plot we sum up the deviations across all categories, to get the so-called total variation distance (TVD). The reported accuracy is then simply reported as 100% - TVD. These accuracies are calculated for all univariate and bivariate distributions. A final accuracy score is then calculated as the average across all of these.

Distances

Identical Matches: 0.0% (0.1%)

Average Distances: 2.19 (1.28)



Explainer

Synthetic data shall be close, but not "too close" to the original data in order to preserve the confidentiality of the original samples. This can be asserted empirically by measuring distances between synthetic records to their closest original records. These distances can then be compared against the observed distances within the original data itself, which serve as a reference for what it means to be "too close". For the visualization above, the distances for the synthetic data are displayed in green, and the distances for the original data displayed in gray. A green line that is significantly left of the gray line within the cumulative density plots would imply that the generated data is too close to the actual records.