Synthetic Data Evaluation Report

This report details the quality, privacy and utility evaluation metrics gained from the synthetic data, and visualisations to help interpret them.

Metrics Summary

Privacy Metrics	Score	Quality Metrics	Score	Utility Metrics	Score
Exact Matches	1.0	Boundary Adherence	1.0	Similarity	0.88
Detection	0.0	Coverage	0.94	Correlation	0.77
Inference Protection	0.19	Complement	0.7	ML Efficacy	0.33
Singling Risk	0.18	nan	nan	nan	nan
Linkability Risk	0.01	nan	nan	nan	nan
Inference Risk	0.07	nan	nan	nan	nan

Synthetic Data Categorisation Level: Correlated Synthetic Data

Correlated Synthetic Data is categorised as the highest risk due to capturing information about relationships and patterns between variables. Therefore, the privacy metrics should be evaluated carefully to ensure individuals arent at risk of being identifiable.

What are the Synthetic Data Categorisation Levels?

Synthetic data exists on a spectrum of low to high fidelity depending on how similar it is to the real data. We can categorise these levels of fidelity as below:

Level 0 (Random Synthetic Data)
Values which are randomly generated, without using the real data or metadata.

Level 1 (Metadata Synthetic Data)
Random data which is generated based on real metadata such as data types, value ranges etc.

Level 2 (Structural Synthetic Data)
Generated by a synthesiser to mimic the structure of the real data, but without capturing the relationships between attributes.

Level 3 (Statistical Synthetic Data)
Roughly keeps the same summary statistics such as mean, median, standard deviation etc.

Level 4 (Correlated Synthetic Data)
Relationships and correlations between variables are captured.

Level 5 (Augmented Data)
Real data is adjusted or nearest neighbours is used to create highly realistic data points which are just slightly different to the real data.

Boundary Adherence Scores

Boundary adherence measures whether values stay within the original min/max ranges of the data. (0.0: means none of the attributes have the same min/max ranges, 1.0: means all attributes have the same min/max ranges)


Top 10 Variables for Boundary Adherence Individual

Bottom 10 Variables for Boundary Adherence Individual
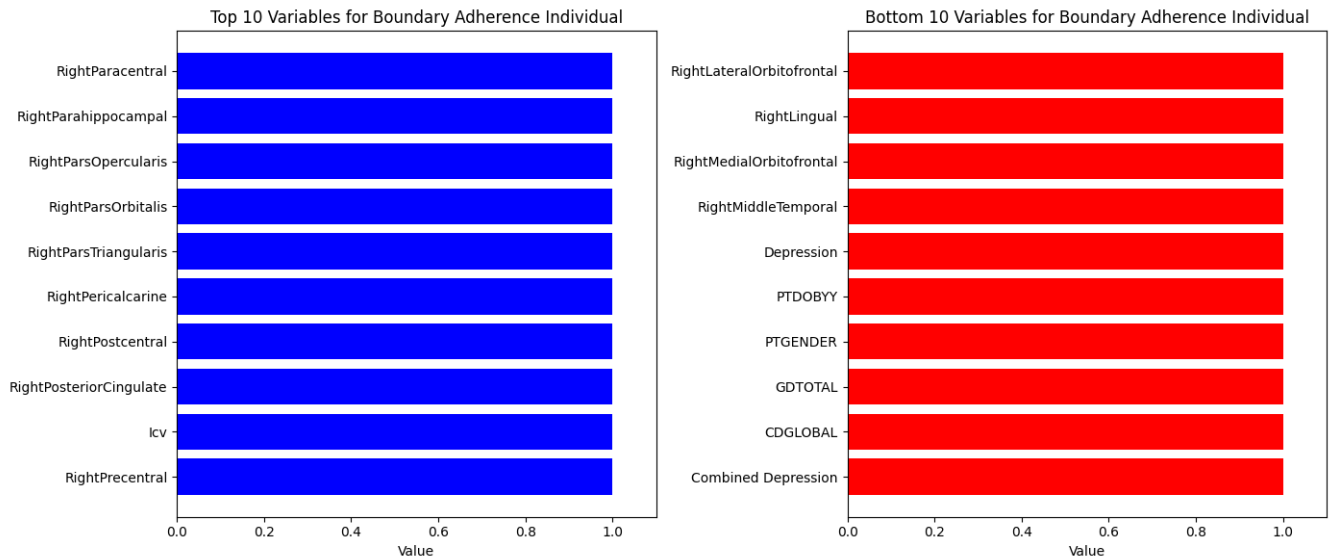
Coverage Scores

Coverage measures whether the whole range of values are represented. (0.0: means none of the values are represented, 1.0: means all values are represented)


Top 10 Variables for Coverage Individual

Bottom 10 Variables for Coverage Individual
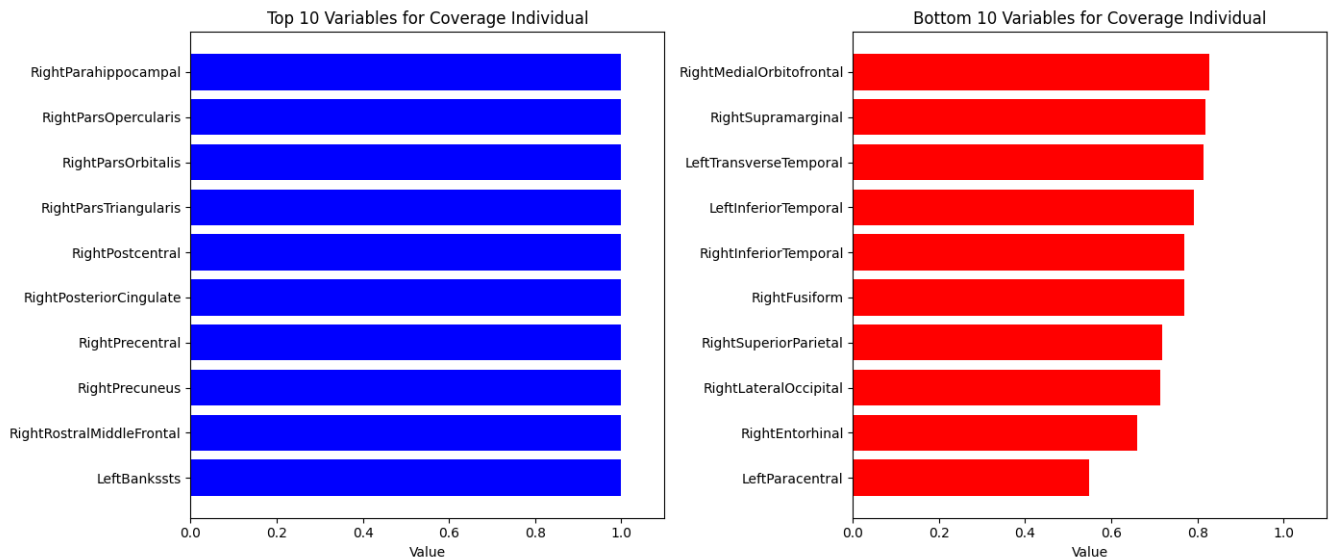
Complement Scores

Complement measures whether the distributions look the same. (0.0: means the distributions are as different as they can be, 1.0: means the distributions are exactly the same)


Top 10 Variables for Complement Individual

Bottom 10 Variables for Complement Individual
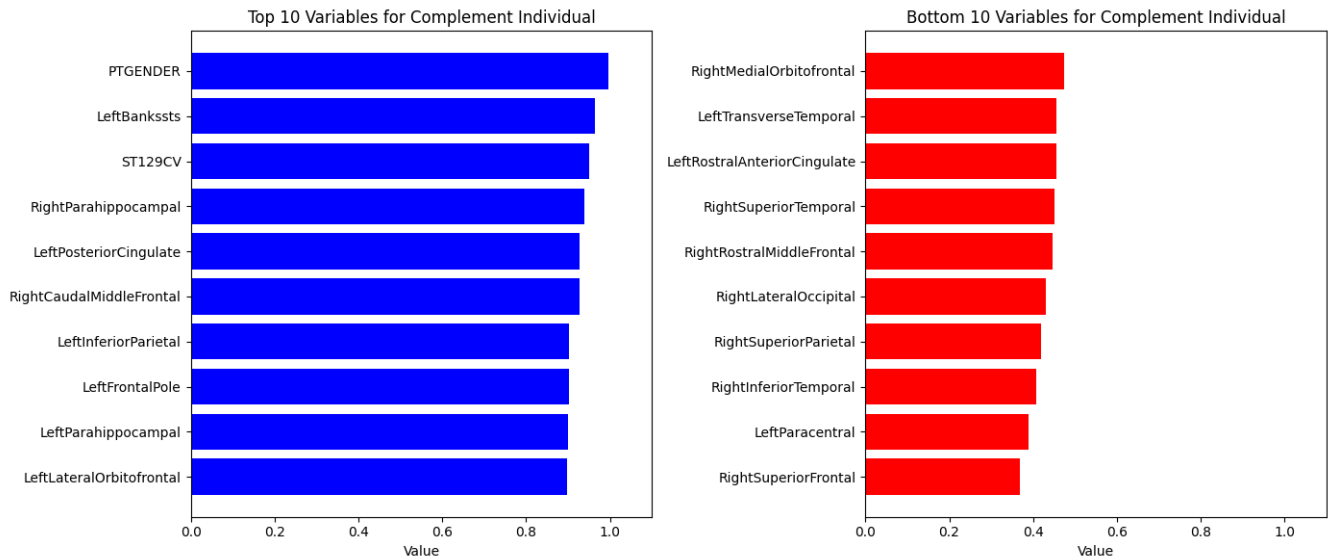
Similarity Scores

Statistic similarity measures how similar the summary statistics are such as mean and standard deviation. (0.0: means the summary statistics are extremely different to each other, 1.0: means the summary statistics are exactly the same)

Top 10 Variables for Similarity Individual

Bottom 10 Variables for Similarity Individual
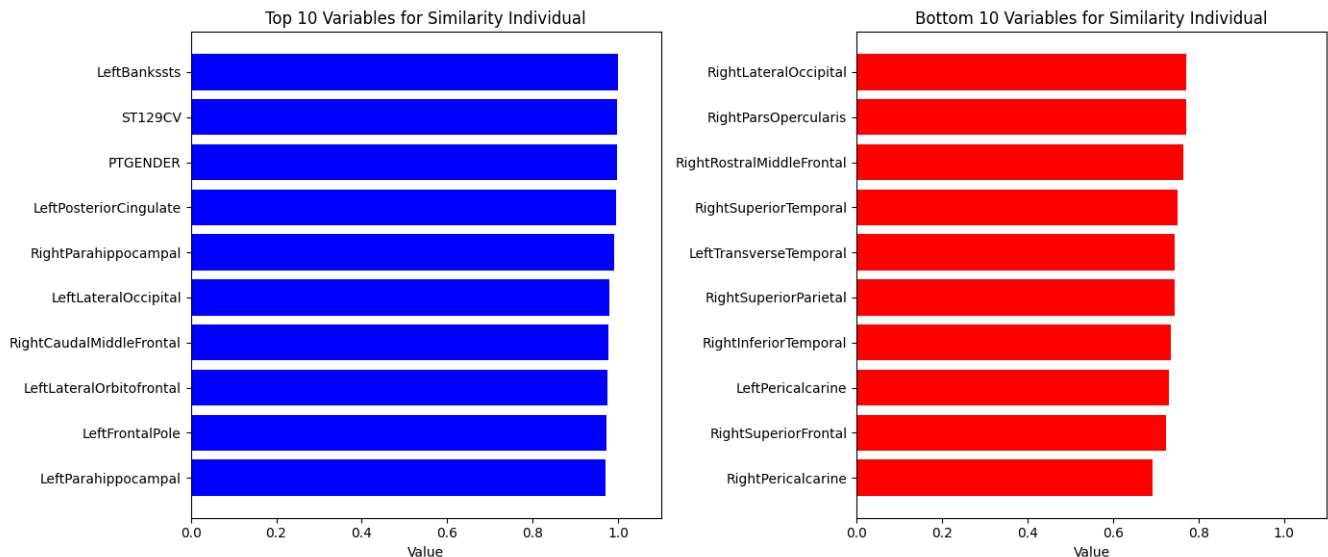
Example Distribution Comparisons
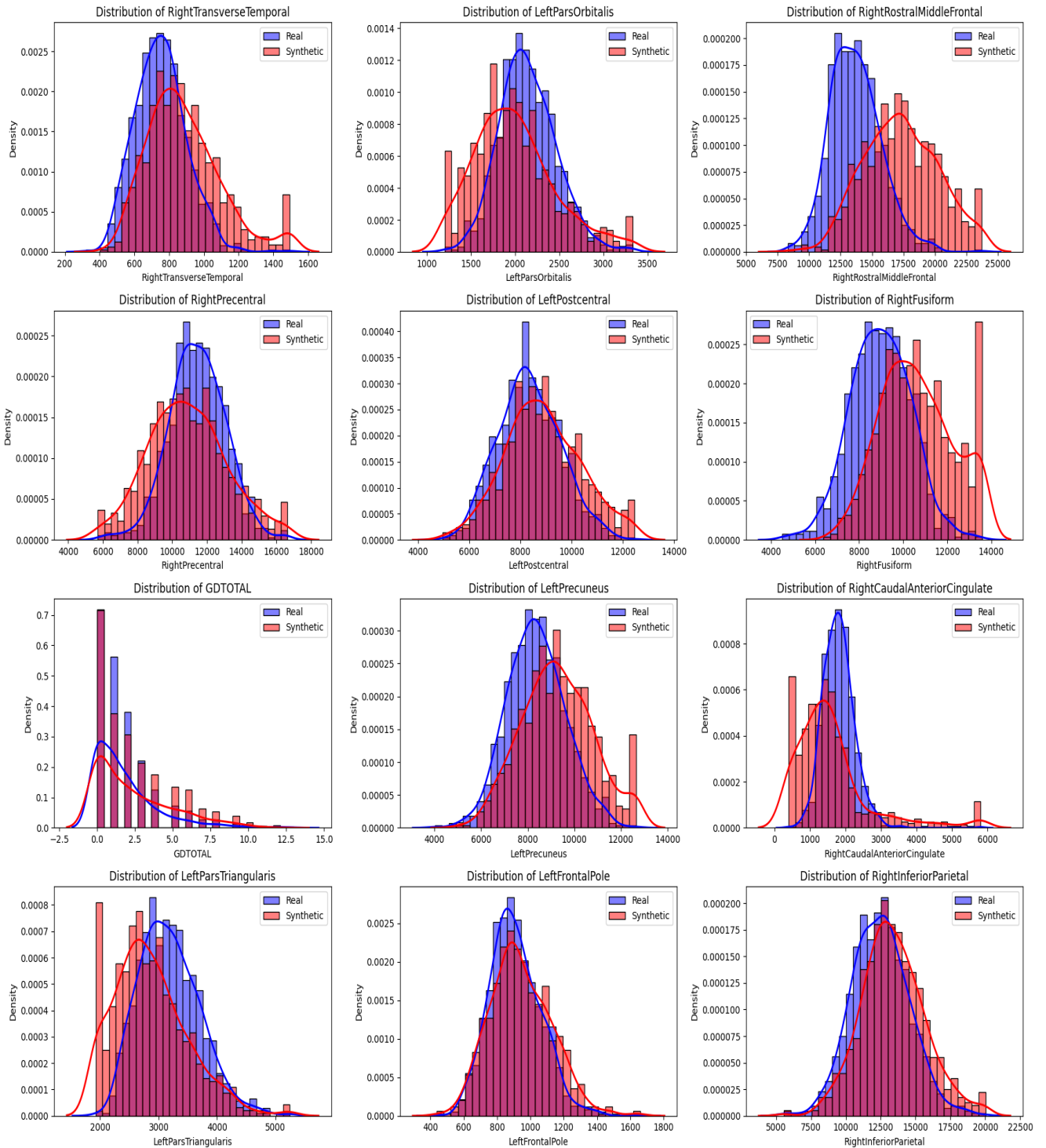
Example Correlation Comparisons
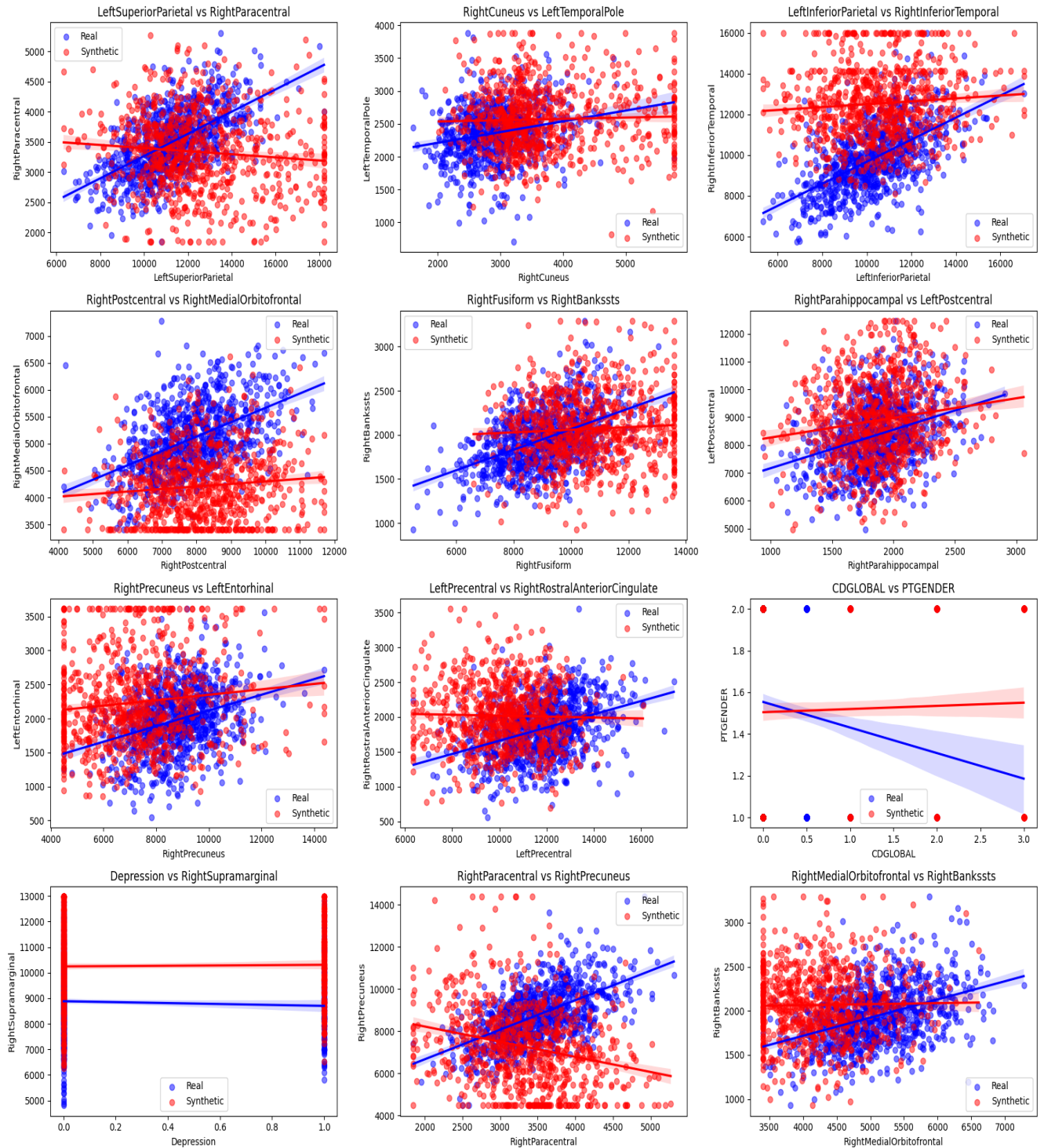
Meaning of Metrics

(Privacy) Exact Matches

This metric measures whether each row in the synthetic data is novel, or whether it exactly matches an original row in the real data.

(best) 1.0: The rows in the synthetic data are all new. There are no matches with the real data.
(worst) 0.0: All the rows in the synthetic data are copies of rows in the real data.

(Privacy) Detection

This metric calculate how difficult it is to tell apart the real data from the synthetic data using machine learning techniques.A score of 1 may indicate high quality but it could also be a clue that the synthetic data is leaking privacy (for example, if the synthetic data is copying the rows in the real data).

(worst) 1.0: The machine learning model cannot identify the synthetic data apart from the real data.
(best) 0.0: The machine learning model can perfectly identify synthetic data apart from the real data.

(Privacy) Inference Protection

This metric calculates the risk of an attacker being able to infer real, sensitive values. It is assumed that an attacker already possess a few columns of real data; they will combine it with the synthetic data to make educated guesses.

(best) 1.0: The real data is 100% safe from the attack. The attacker is not able to correctly guess any of the sensitive values by applying the chosen attack algorithm.
(worst) 0.0: The real data is not at all safe from the attack. The attacker is able to correctly guess every sensitive value by applying the chosen attack algorithm.