SYNTHETIC DATA QUALITY REPORT

268 # REAL DATA RECORDS

268 # SYNTHETIC DATA RECORDS GENERATED

7 # COLUMNS

91%

FIDELITY SCORE

72%

UTILITY SCORE

73%

PRIVACY SCORE

FIDELITY SCORES

Fidelity measures how well the synthetic data statistically matches the original records. It is provided through univariate and multivariate metrics, model and assumption-free.

STATISTICS DISTANCE

The metrics below provide summarized information of the detailed statistical metrics calculated through YData's profiling.

0.98

CORRELATION SIMILARITY

0.96

DISTANCE DISTRIBUTION

The **CORRELATION SIMILARITY** measures how close are synthetic and real correlation matrices. It is bounded between [0-1] metric, the closer to 1, the higher fidelity.
DISTANCE DISTRIBUTION measures the features' distribution similarity between original and generated data. The Chi-squared test evaluates features with discrete distributions, and the Kolmogorov-Smirnov test evaluates features with continuous distributions. Returns values between [0, 1], closer to 1 is desirable.

MISSING VALUES SIMILARITY

The **MISSING VALUES SIMILARITY** measures how close are the percentages of missing values in the synthetic and real data. This metric is bounded between [0-1], where 1 represents the same percentage of missing data.

1.00

MISSING VALUES SIMILARITY

STATISTICAL SIMILARITY

The **STATISTICAL SIMILARITY** measures how similar are the synthetic and real data considering five metrics: mean, standard deviation, median, 25% quantile, and 75% quantile. Each similarity is bounded between [0-1], where 1 represents equal values. Only numerical features are considered in this analysis.

0.97

MEAN

0.99

STD. DEV.

0.97

MEDIAN

0.98

Q25%

0.96

Q75%

Feature	Mean	Std. Dev.	Median	Q25%	Q75%
age	0.95	1.00	0.93	0.95	0.93
bmi	1.00	0.98	1.00	1.00	0.99
charges	0.98	1.00	0.97	0.98	0.94

DISTRIBUTION METRICS

The distribution metrics compare the probability distributions of the real and synthetic data variables. These metrics return values between [0, 1], where closer to 1 is desirable (i.e., the distributions are likely the same).

KOLMOGOROV-SMIRNOV TEST

The **KOLMOGOROV-SMIRNOV (KS) TEST** compares the distribution between two continuous variables (real and synthetic data) using the empirical CDF. The two tables below present the five features with the highest and lowest values for this test.

0.42

KOLMOGOROV-SMIRNOV TEST

Feature	KS Test (Highest)
bmi	0.80
age	0.28
charges	0.19

Feature	KS Test (Lowest)
bmi	0.80
age	0.28
charges	0.19

CHI-SQUARED TEST

The **CHI-SQUARED (χ2) TEST** compares the distribution between two categorical variables (real and synthetic data). The two tables below present the five features with the highest and lowest values for this test.

0.99

CHI-SQUARED TEST

Feature	χ2 Test (Highest)
children	1.00
region	1.00
smoker	0.99
sex	0.98

Feature	χ2 Test (Lowest)
children	1.00
region	1.00
smoker	0.99
sex	0.98

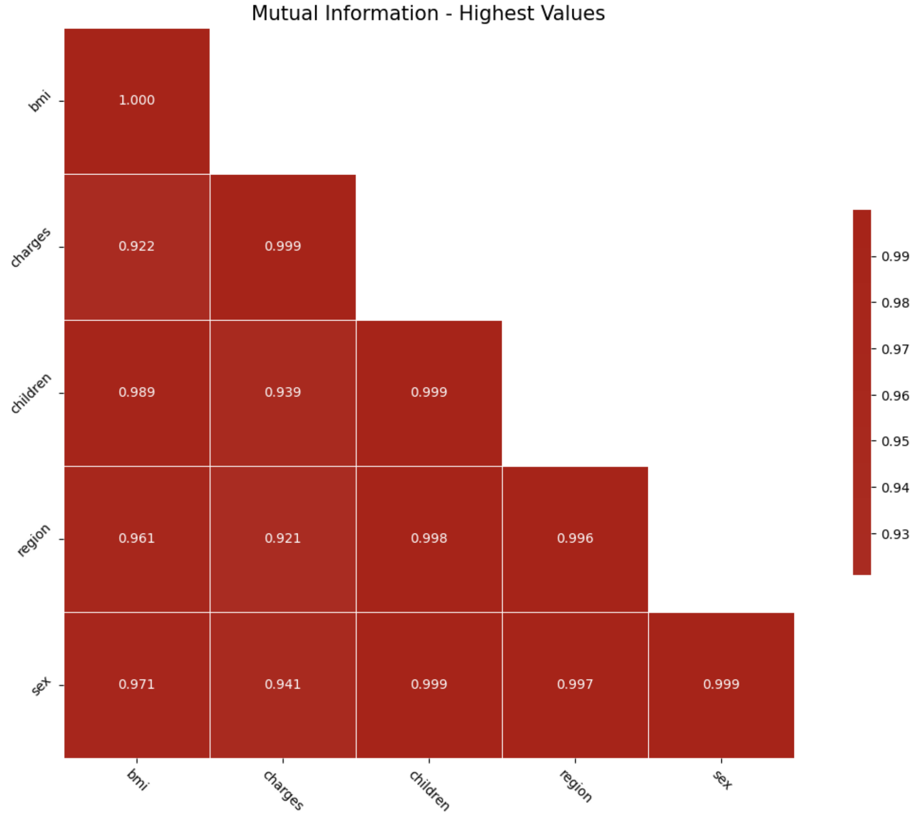
MUTUAL INFORMATION

MUTUAL INFORMATION (MI) measures how much information can be obtained about one feature by observing another. This metric calculates the similarity between real and synthetic MI values for each pair of features. It returns values between [0, 1], where closer to 1 is desirable (i.e., equal MI).

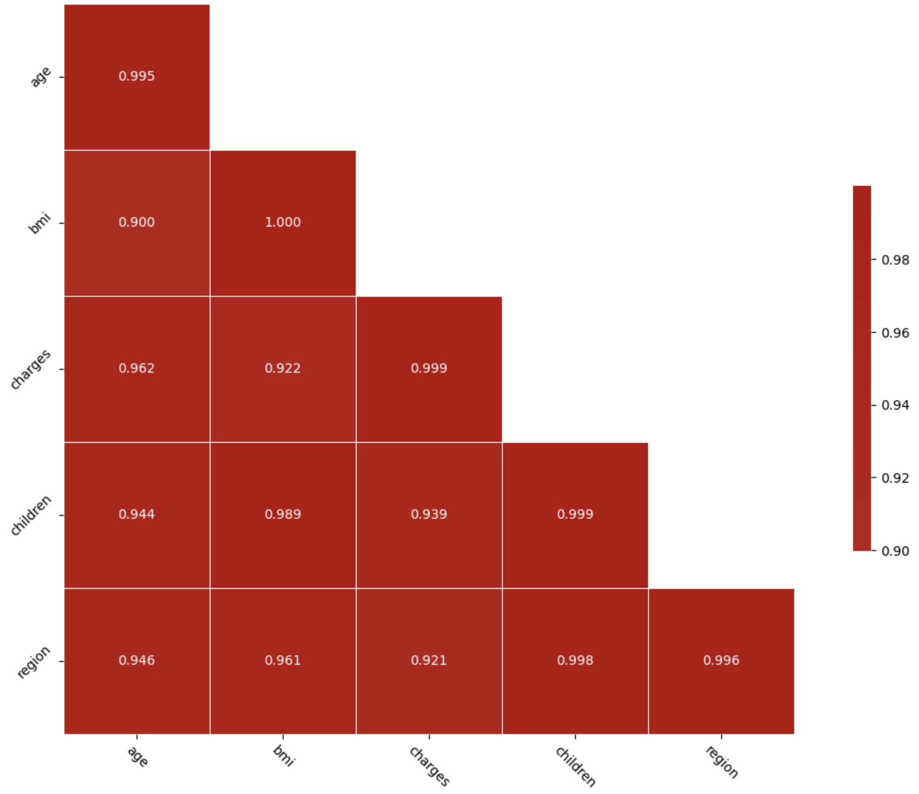
0.97

MUTUAL INFORMATION

Mutual Information - Highest Values



Mutual Information - Lowest Values



COVERAGE METRICS

The coverage metrics describe how well the real data variables are represented in the synthetic data and they return values between [0, 1], where 1 represents complete coverage. These metrics are divided into two groups:

- Metrics for categorical data, which include **Category Coverage** and **Missing Category Coverage**.
- Metrics for numerical data, namely the **Range Coverage**.

CATEGORICAL DATA

The following metrics are specific to categorical data.

CATEGORY COVERAGE

The **CATEGORY COVERAGE (CC)** computes the ratio of real data categories that are represented in the synthetic data. The two tables below present the five features with the highest and lowest coverage.

1.00

CATEGORY COVERAGE

Feature	CC (Highest)
sex	1.0
children	1.0
smoker	1.0
region	1.0

Feature	CC (Lowest)
sex	1.0
children	1.0
smoker	1.0
region	1.0

MISSING CATEGORY COVERAGE

The **MISSING CATEGORY COVERAGE (MCC)** computes the similarity ratio of categorical values from the real data that are not represented in the synthetic data. The two tables below present the five features with the highest and lowest coverage.

0.56

MISSING CATEGORY COVERAGE

Feature	MCC (Highest)
region	0.75
sex	0.50
children	0.50
smoker	0.50

Feature	MCC (Lowest)
region	0.75
sex	0.50
children	0.50
smoker	0.50

NUMERICAL DATA

The following metrics are specific to numerical data.

RANGE COVERAGE

The **RANGE COVERAGE (RC)** computes the similarity ratio between the numerical variables domain in the real dataset compared to the synthetic one. The two tables below present the five features with the highest and lowest coverage.

0.98

RANGE COVERAGE

Feature	RC (Highest)
age	1.00
bmi	1.00
charges	0.94

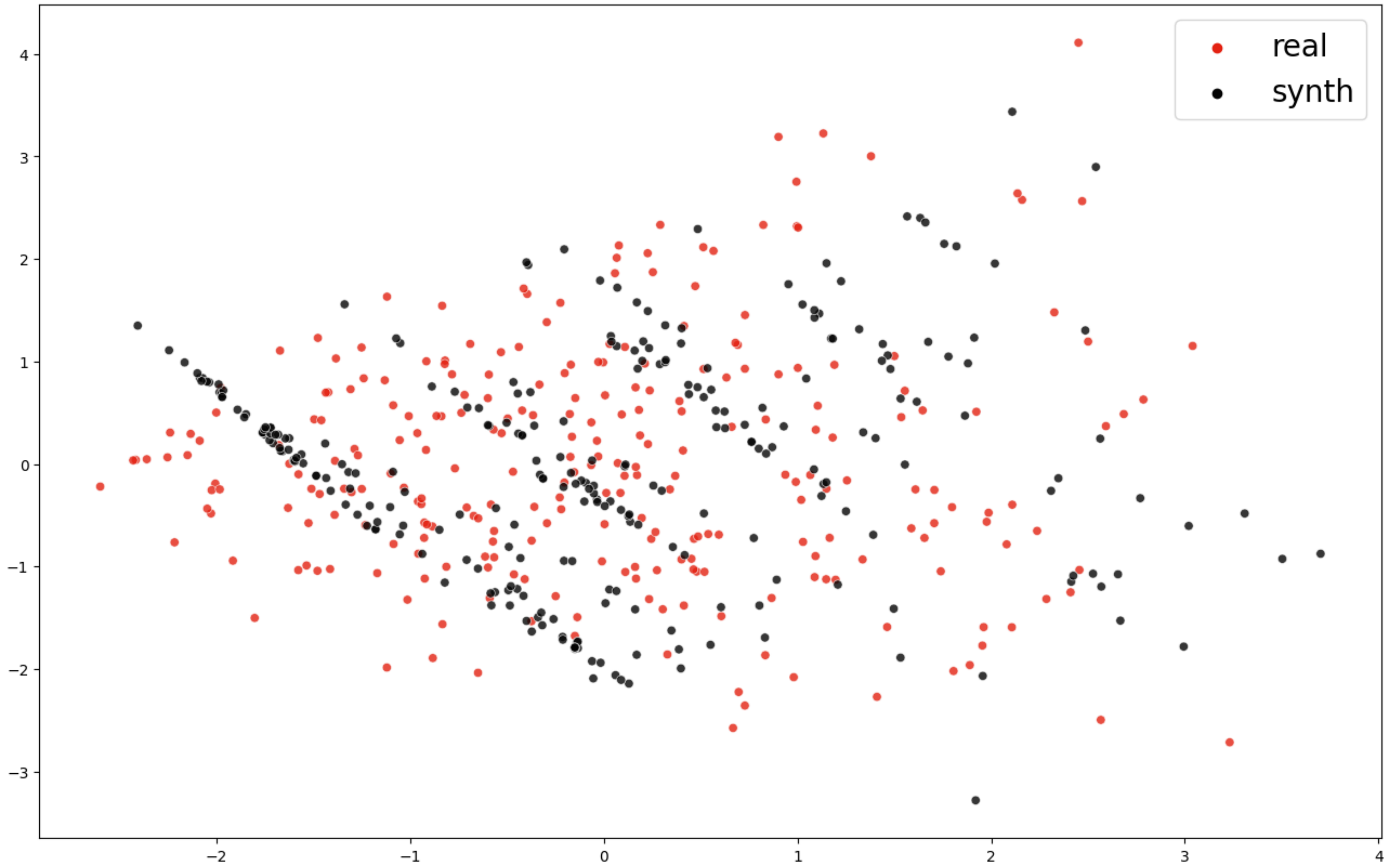
Feature	RC (Lowest)
age	1.00
bmi	1.00
charges	0.94

DIMENSIONALITY REDUCTION

The dimensionality reduction visualization plots show how closely the distribution of the synthetic data resembles that of the original data on a two-dimensional graph. Principal Component Analysis (PCA) algorithm used to reduce the datasets dimensionality.

PCA captures any fundamental difference in the distributions of the datasets. The scatterplots represent depict this difference visually.

Ate represent the two first main Eigenvectors that together explain 45.81% of of the total variance of the dataset.



UTILITY SCORES

YData's Synthetic Data Profile utility scores are grouped into two main classes:
• Query quality enables you to understand the quality of the generated synthetic data to answer the questions with similar results as the original data.
• The Predictive performance allows you to understand the effects of an ML model trained on your synthetic data and later applied in production applications. It provides a supervised perspective of the synthetic data generated.

QSCORE

0.72

QSCORE

QScore measures the utility of the synthetic data by comparing the returns of random aggregation-based queries on both the synthetic and original datasets. The QScore ensures that future queries performed on the synthetic data would return the same statistical characteristics as those on the original data. Score between [0, 1]. It is recommended a QScore above 0.8 if leveraging the synthetic data for BI activities or unsupervised learning.

PRIVACY SCORES

Privacy indicates the level of confidentiality of the synthetic data. Larger and more complex datasets typically offer higher privacy scores and more protection against attacks.

0.0	**1.0**	**0.07**	**0.84**
EXACT MATCHES	MEMBERSHIP INFERENCE SCORE	NEIGHBOURS PRIVACY	SYNTH CLASSIFIER

The **EXACT MATCHES** score counts the percentage of sensitive records in the synthetic data that match the records in the original dataset. It is bounded between [0-1]. The score must be 0 for safe data-sharing.
The **MEMBERSHIP INFERENCE SCORE** score measures the risk that an attacker can determine whether a particular record of the original dataset was used to train the synthesizer. A membership inference score close to 1 indicates that an attacker is unlikely to determine if a specific record was a member of the original dataset used to train the synthesizer. This metric is bounded between [0,1].
The **NEIGHBOURS PRIVACY** score quantifies the risk of synthetic data points lying too close to the original data points. The score is calculated by evaluating the results of a high-dimensional nearest-neighbors search on the synthetic data overlapped with the original data. A score close to 0 indicates that the synthetic points may lie too close to the original data points, and it is not safe to share.
The **SYNTHETIC CLASSIFIER** The synthetic classifier score, translated the ROC-AUC of a model trained to distinguish between real and synthetic data. A score close to 1 indicates that the estimator is not able to discriminate between the original and synthetic data records, making the data safer to be shared. This metric is bounded between [0,1].