SYNDATA

Evaluation report of synthetic dataset

discharged_data_train.csv

Introduction 2
Model 1 2
Features similarity 2
a. Individual distributions 3
Wasserstein Distance 3
Individual distributions visuals 3
Overall WD distribution 5
b. Pairwise distributions 5
Pearson's correlation 5
Mutual Information 6
Missing Values 7
a. Missing Values Matrices 7
Pearson's Correlation 8
Mutual Information 8
Risk of Re-Identification 9
Appendix 1: Model 2 9
Features similarity 10
Missing Values 15
Risk of Re-Identification 16
Appendix 2: Model 3 17
Features similarity 17
Missing Values 22
Risk of Re-Identification 23
Models' comparison 24

Introduction

The evaluation reports assess the quality of the synthetic datasets produced by Syndata based on the original datasets provided by Region Västerbotten (RVB). Syndata produced 2 evaluation reports: 1 evaluation report covering model 1 and 2 for *recovered* dataset and 1 report for *discharged* dataset.

After multiple test runs, Syndata concluded with 2 best models. These 2 models can be used by RVB to sample synthetic datasets of their preferred size. Syndata concludes that both models have the potential for quality synthetisation. Given good quality synthetic datasets, RVB can achieve its goal of predicting patients' recovery.

The current report evaluates the characteristic of the synthesized **discharged_data_train.csv** dataset sampled with **Model 1**. This report considered general statistics and visuals as comparison tools between the original and synthesized dataset. The size of the synthetic datasets evaluated are of the same size as the original (eg. "1x").

Comparison datasets:

- Discharge_data_train.csv
- Discharged_synthetic_model1__1x_dataset.csv

The datasets and the evaluation frameworks (as jupyter notebooks) are available on the RVB server.

Characteristics of the original *discharged* dataset:

- contains both categorical values and non-categorical
- 7 fields with repeated information
- 105 fields(features/columns) and 1689 observations (data points/rows)

Model 1

Model 1 uses CTGAN networks, a collection of Deep Learning based Synthetic Data Generators for single table data, which are able to learn from real data and generate synthetic clones with high fidelity. The CTGAN model is available in **sdv** library.

Features similarity

Once we have created a synthesized datasets of the same size as the original, the next step is to visualize how well the properties of each feature have been preserved. A first method is to evaluate the individual distributions one by one. As a secondary method we will be looking at pairwise distributions to understand how well the relations between features are preserved in the synthetic datasets.

a. Individual distributions

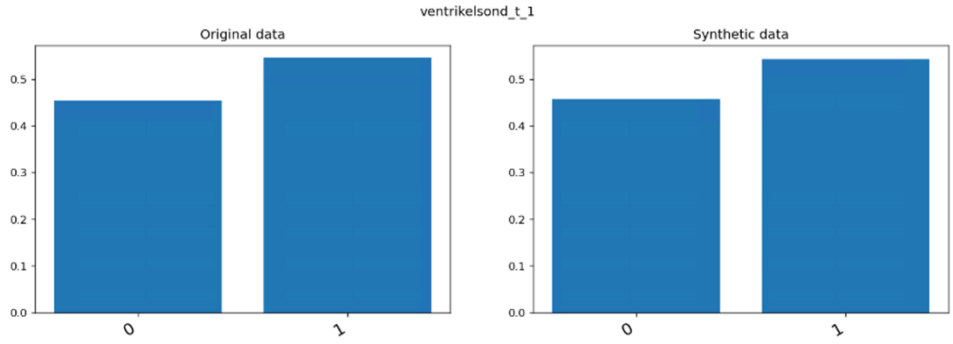
Wasserstein Distance

We compare individual distributions one-to-one using a distance metric called Wasserstein Distance (WD). A value of 0 means that the two individual distributions are identical.

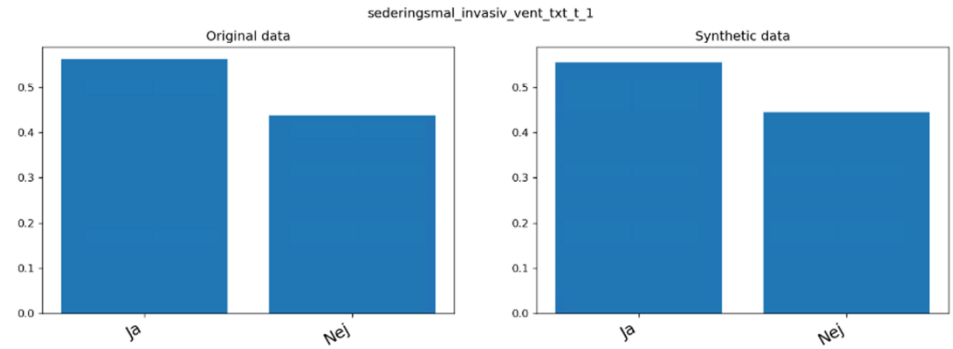
This distance is also known as the earth mover's distance, since it can be seen as the minimum amount of "work" required to transform one distribution into another one, where "work" is measured as the amount of distribution weight that must be moved, multiplied by the distance it has to be moved.

Individual distributions visuals

Examples of similar distributions



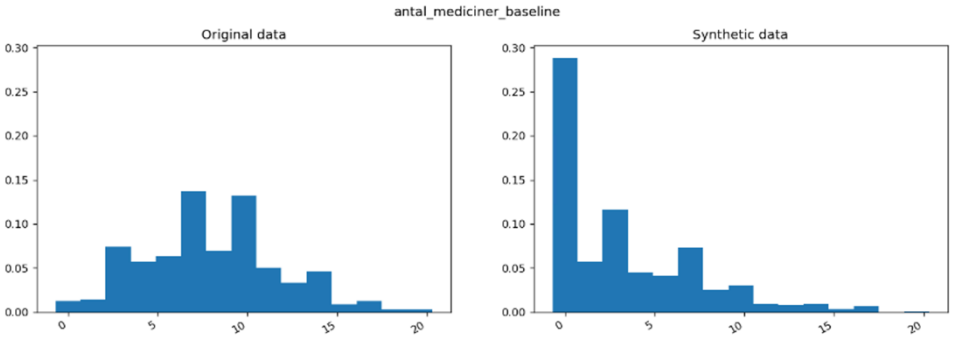
WD: 0.002



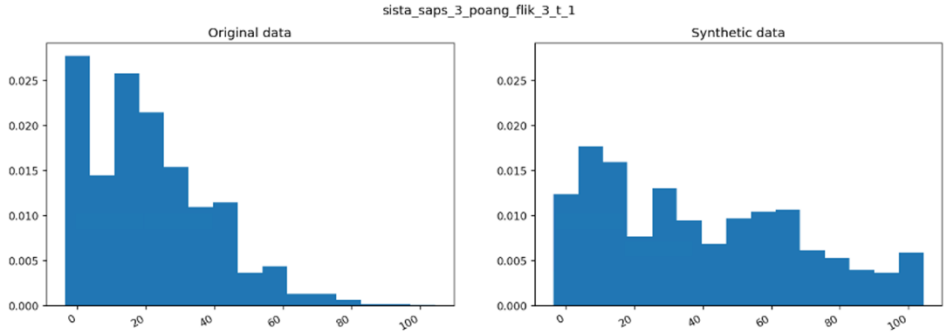
WD: 0.0071

Interpretation: The most similar 5 features are categorical features such as *noradrenalin_baseline, cvk_1_t_1, diff01_sederingsmal_invasiv_vent_cat, antal_mediciner_t_1, sedering_cat_t_1*. Distributions have WD close to 0 indicating a close replica to the original.

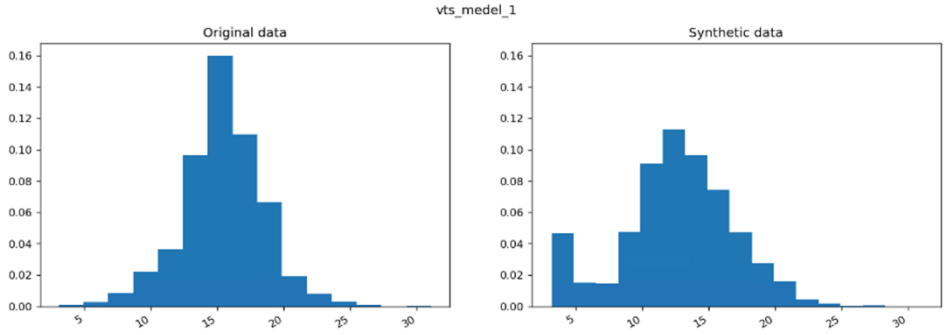
Examples of less similar distributions:



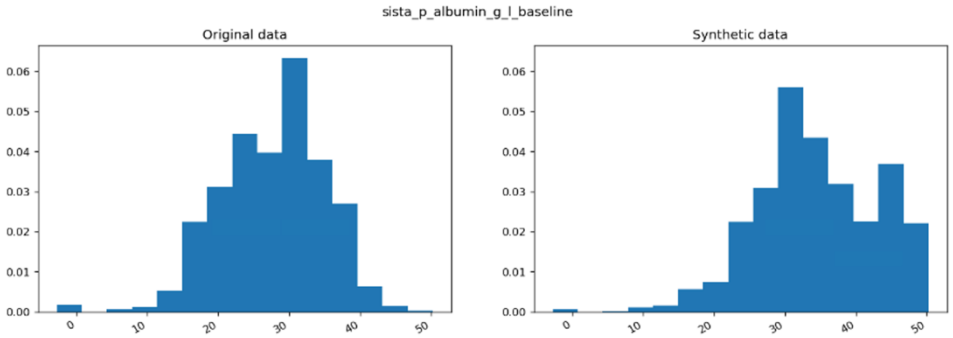
WD: 1.29



WD: 0.99



WD: 0.95



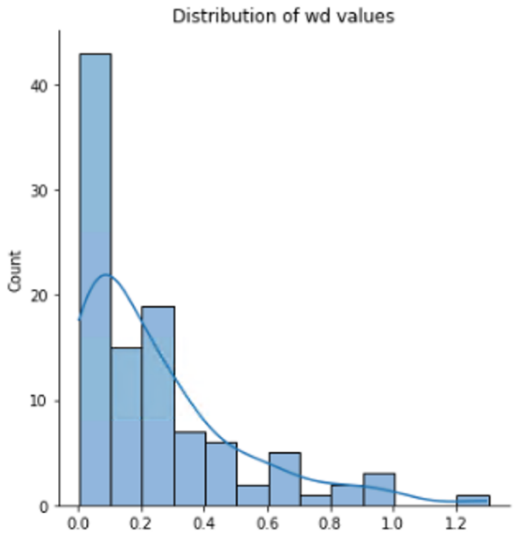
WD: 0.93

Interpretation: The least similar 4 fields distributions have WD below 1, except one outlier. The least similar features are non-categorical. However, the distribution shape is similar even though the high WD distance.

Overall WD distribution

As a way to summarize the individual distributions quality, we use a distribution plot of all Wasserstein Distances. The closer the mean to 0 and the smaller the standard deviation, the more similar features are to the original dataset.

```
General statistics of WD distribution:

WD Distribution Mean: 0.2431
WD Distribution Standard Deviation: 0.2581
```



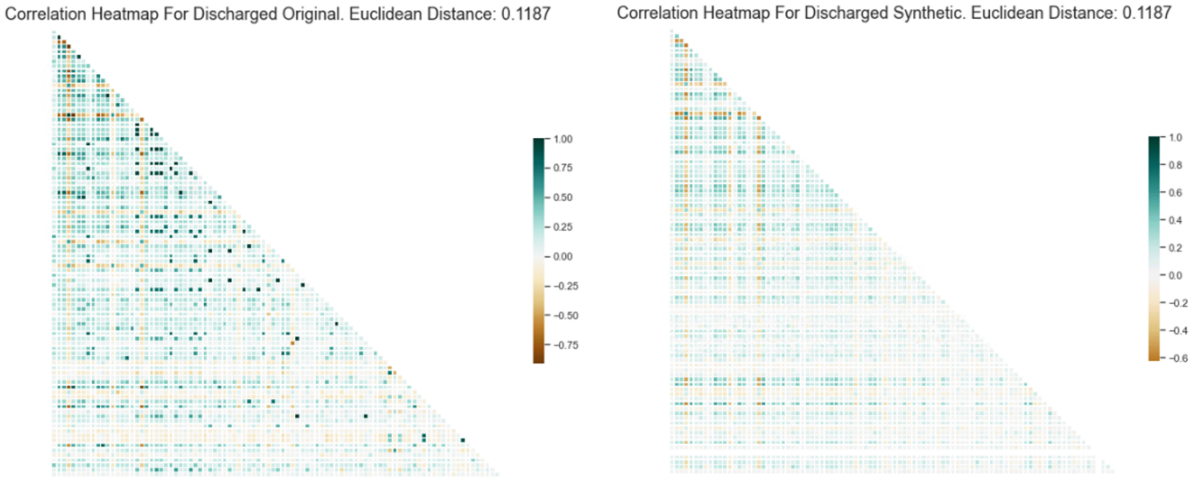
Interpretation: Most of the fields have distributions with WD distances below 0.5 which indicates strong similarity of the synthetic distributions to the original. There is one outliner, namely *antal_mediciner_baseline,* a non-categorical feature.

b. Pairwise distributions

Pearson's correlation

Pearson's correlation coefficient measures the strength of a linear association between two variables. Its value range spans over [-1,1]. A value of 0 indicates that there is no association between the two variables. We compare the distance between these 2 matrices by the Euclidean distance. A point with an increased colour intensity means a pair of 2 features have high correlations.

Finally, we compare the distance between these 2 matrices with an Euclidean distance. The Euclidean Distance for the correlation matrix ranges on a **[0, 2] scale.**



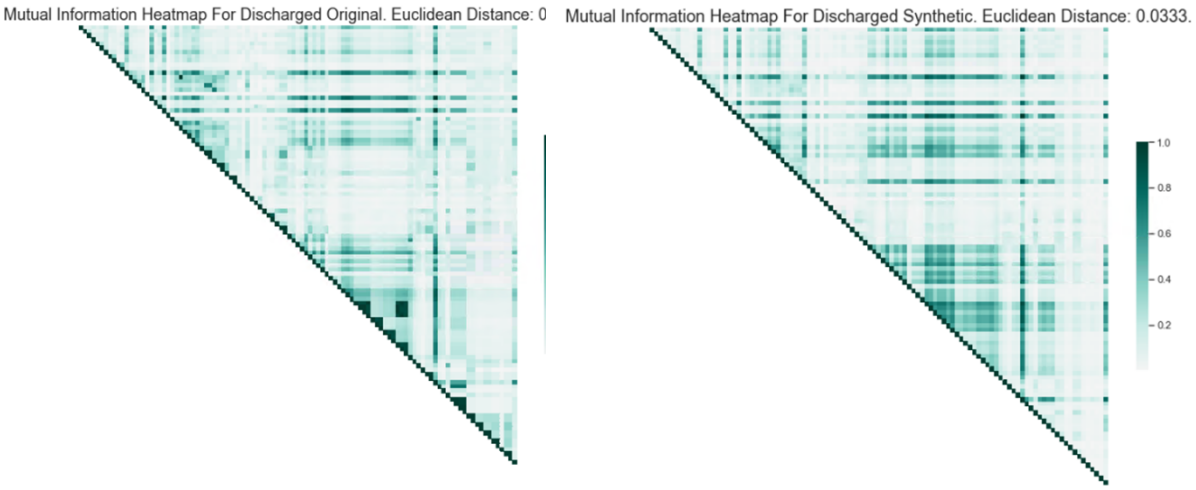
Euclidean distance: 0.1187

Interpretation: Overall patterns of correlations are preserved. Some strong positive correlations show a lower intensity in the synthetic dataset. The overall distance is small indicating a high correlation of pairwise distributions.

Mutual Information

Mutual information measures the relatedness between two random variables. It takes values on a [0,1] interval. A value of 0 (associated with a lower colour intensity) indicates that there is no relatedness between the two variables. We use the Mutual Information metric to plot a heatmap which visualises the level of relatedness between any 2 features. A higher intensity of the point indicates a higher relatedness between 2 specific features.

Finally, we compare the distance between these 2 matrices with a Euclidean distance. The Euclidean Distance for the mutual information matrix ranges on a **[0, 1] scale.**



Euclidean distance: 0.0333

Interpretation: General vertical and horizontal patterns are preserved. Synthetic dataset loses some of the sharp details. However, the overall distance is small indicating missing value patterns are preserved between datasets.

Missing Values

We evaluate the relation between missing values with Mutual Information (to capture non-linear relations) and Pearson's Correlation (to capture linear relations).

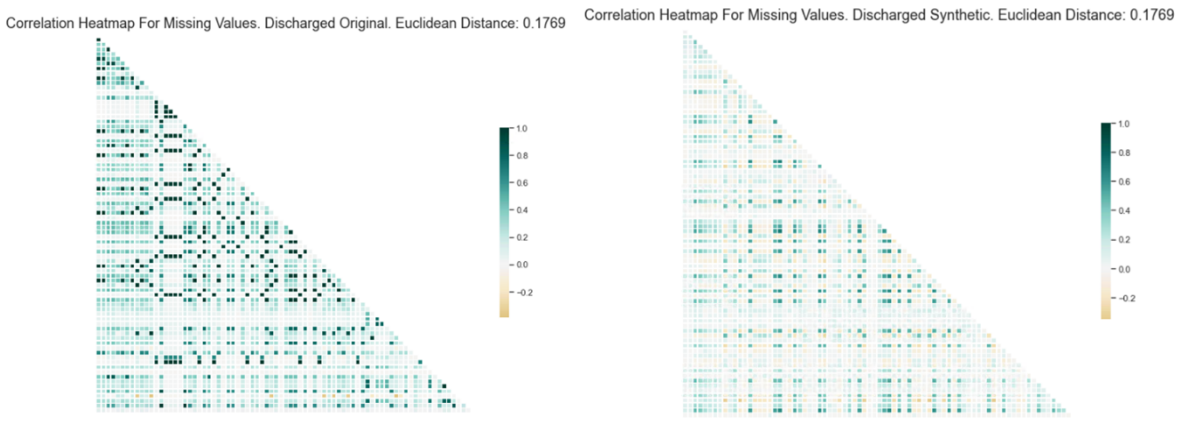
Euclidean Distance is used to measure the similarity between the 2 matrices. It compares the differences between two correlation (or mutual information) matrices. The closest to 0, the more similar the 2 matrices are against each other.

We have a different distance scale for each metric. The Euclidean Distance for the correlation matrix ranges on a **[0, 2]** scale, while the distance for the mutual information matrix ranges on a **[0, 1]** scale.

a. Missing Values Matrices

Pre-steps: We map the original and synthetic datasets values to matrices that indicate null or not null values. Later, we remove the columns without any variation. Finally, we compute the correlation /mutual information matrix.

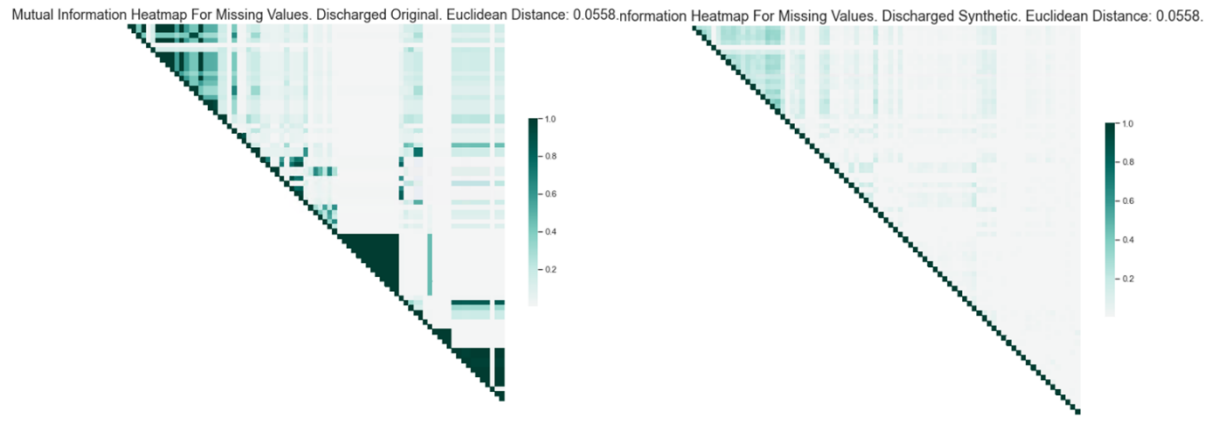
Pearson's Correlation



Euclidean distance: 0.176

Interpretation: General patterns are preserved. Synthetic dataset loses some of the sharp details, especially some of the positive correlations. Overall distance is small indicating missing value patterns are preserved between datasets.

Mutual Information



Euclidean distance: 0.0558

Interpretation: Upper corner preserves the patterns, but at a lower intensity. Some details are lost more than desired, but the distance is small indicating that the missing values relatedness is kept in

the synthetic dataset. The faded colours can be attributed to the small missing value rate present in the original dataset which is entirely replaced in the synthetic.

Risk of Re-Identification

Risk of re-identification is a relative metric to the original dataset characteristics. It indicates the minimum and maximum risk of the individuals from the synthetic dataset to be re-identified.

The score is influenced by number of: (1) unique values in each column (2) number of identical values per individual with any point in the original dataset.

Statistic	Measurement
RIR min	0.24
RIR max	0.52
RIR mean	0.38
RIR scale	[0,2]

Interpretation: A synthetic individual with a risk of 0 means that there is no data point in the original dataset that has any same values. The highest risk of a synthetic datapoint is 0.52, in the bottom 26% overall risk. To reduce the overall risk, the individuals above a certain risk threshold can be eliminated from the synthetic dataset.

Appendix 1: Model 2

Model 2 is based on copula functions, available as part of **sdv** library.

Comparison datasets:

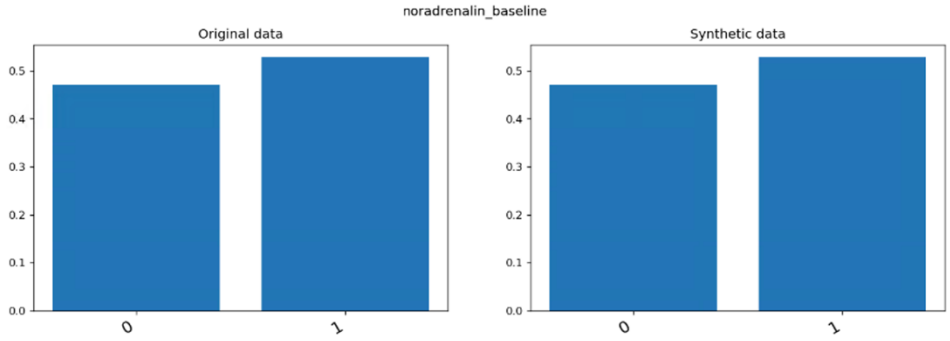
- Discharge_data_train.csv
- Discharged_synthetic_model1__1x_dataset.csv

Features similarity

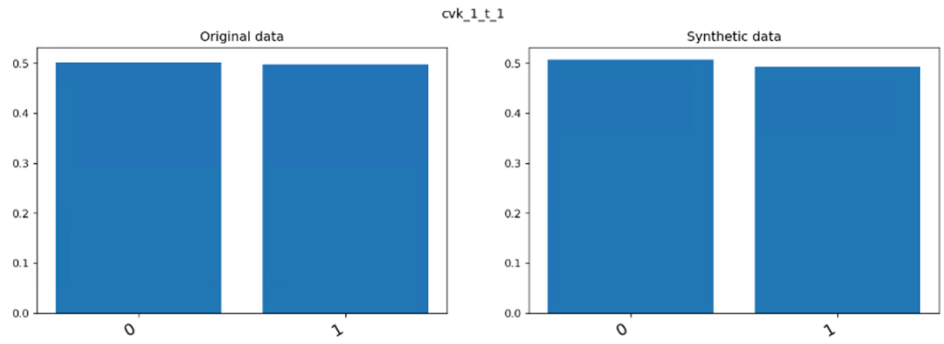
a. Individual distributions

Individual distributions visuals

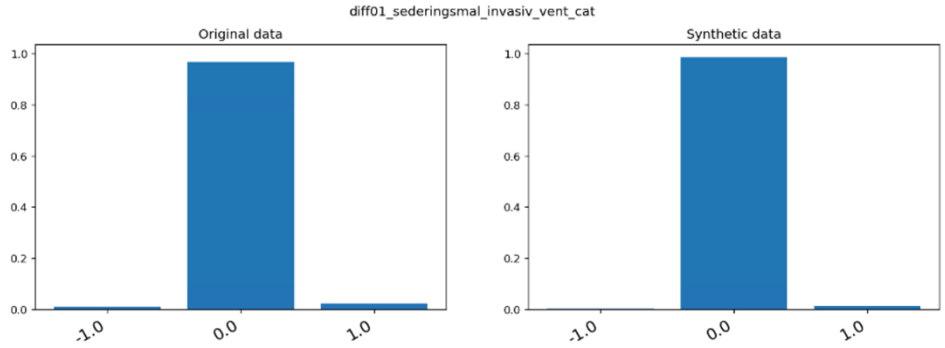
Examples of similar distributions



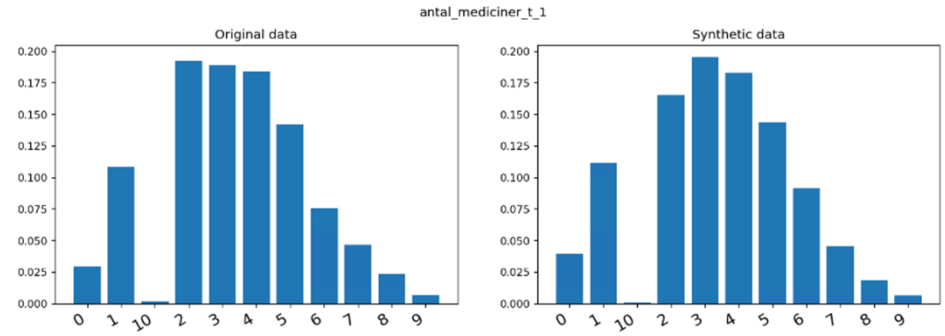
WD: 0.0003



WD: 0.005



WD: 0.009

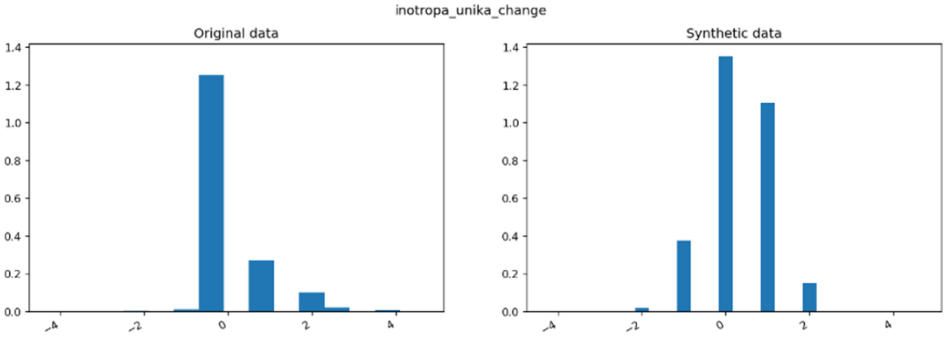


WD: 0.0091

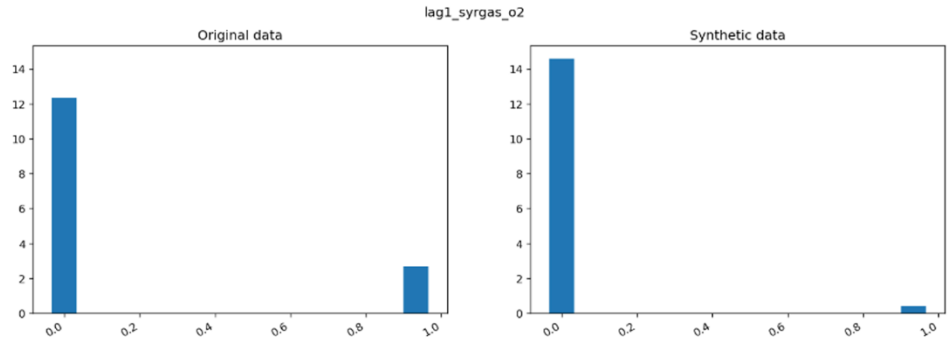
Interpretation: The most similar 5 features are categorical features with a WD close to 0 indicating a close replica to the original. These include *noradrenalin_baseline, cvk_1_t_1 , diff01_sederingsmal_invasiv_vent_cat, antal_mediciner_t_1 sedering_cat_t_1.*

The same features are also found in the top features synthesised with model 2. Thus, both models capture well the distributions of these categorical values.

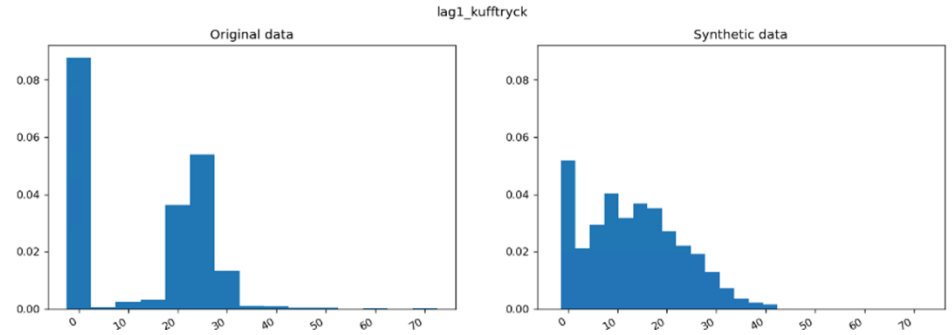
Examples of less similar distributions:



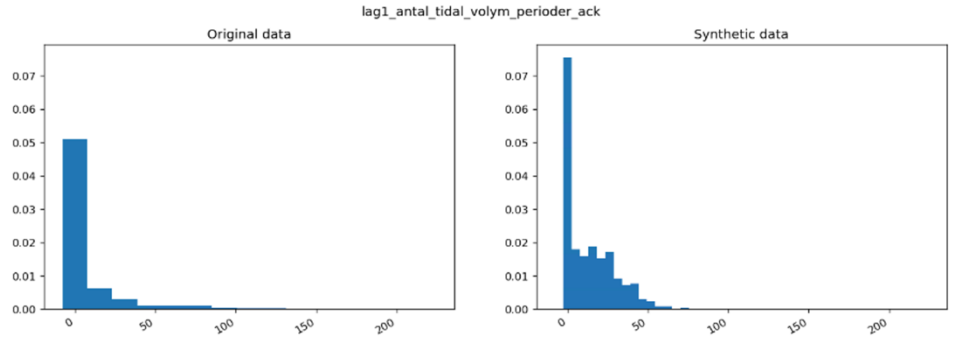
WD: 0.495



WD: 0.395



WD: 0.3461

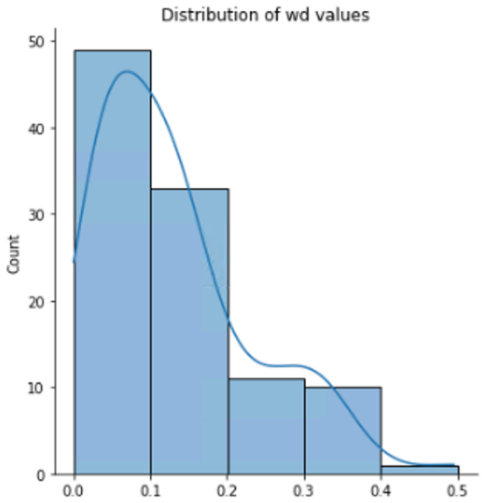


WD: 0.343

Interpretation: The least similar 4 fields display WD below 0.49. Model 2 is able to improve the WD of distributions such as *Antal_mediciner_baseline, sista_saps_3_poang_flik_3_t_!, vts_medel_1, sista_p_albumin_d_l_baseline.* The bottom worse distributions are not preserved between the models.

Overall WD distribution

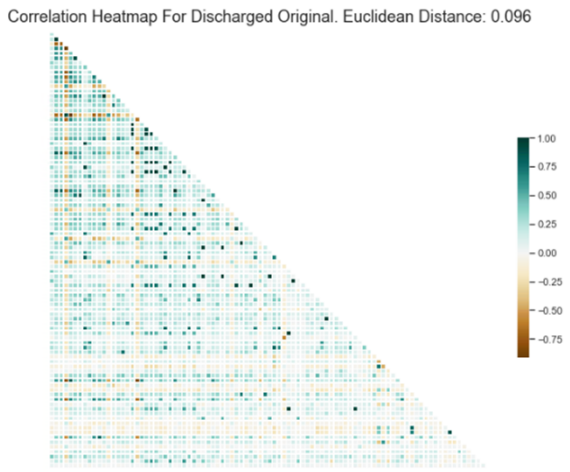
```
General statistics of WD distribution:

WD Distribution Mean: 0.1307
WD Distribution Standard Deviation: 0.1021
```

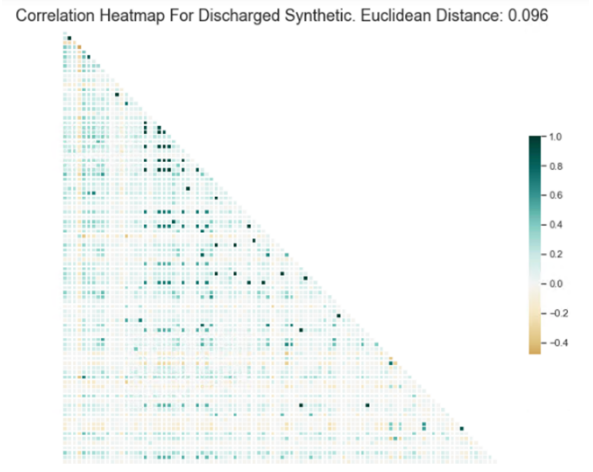


Interpretation: As model 2 was able to correct some of the least similar distributions, the overall WD upper bound was lowered to 0.49. By assessing the mean and standard deviation above, we can conclude that model 2 is able to produce better individual distributions.

b. Pairwise distributions

Pearson's correlation

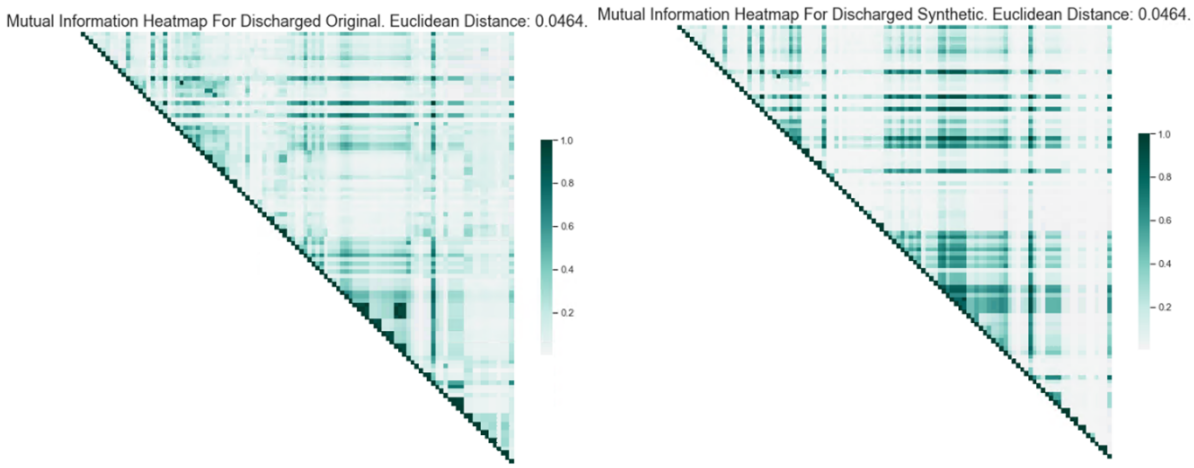




Euclidean distance: 0.099

Interpretation: Overall patterns are preserved. Model 2 shows an improvement in preserving some of the strong positive correlations leading to an improved Euclidean distance from **0.1187** (model 1) to **0.099** (model 2). Overall, Model 2 captures better the correlation of fields.

Mutual Information



Euclidean distance: 0.0464

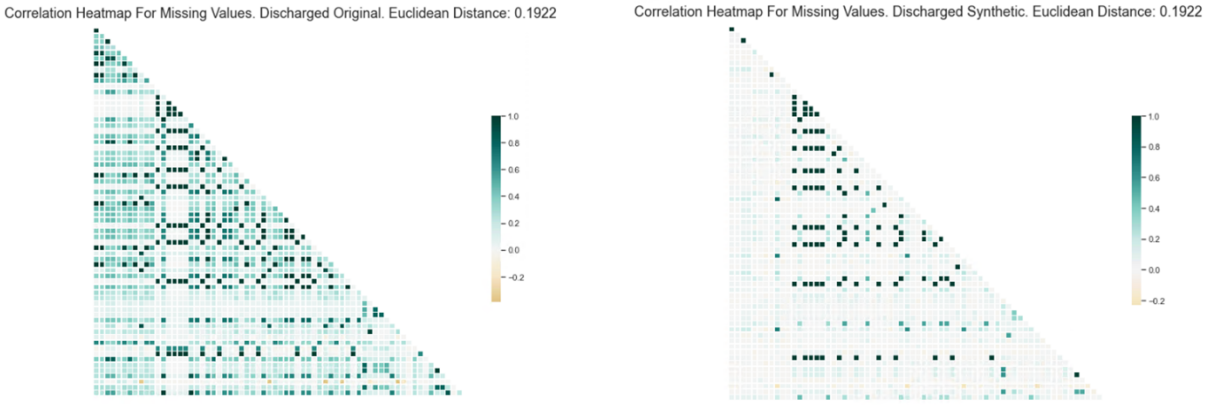
Interpretation: Pairwise distribution patterns are preserved between original and synthetic. Some of the strong mutual information values are preserved well. The overall distance is small indicating a high relatedness between fields.

The Euclidean distance increased from **0.033** (model 1) to **0.046** (model 2). Thus, model 1 outperforms model 2 by producing relatedness, however, the difference is small.

Missing Values

a. Missing Values Matrices

Pearson's Correlation

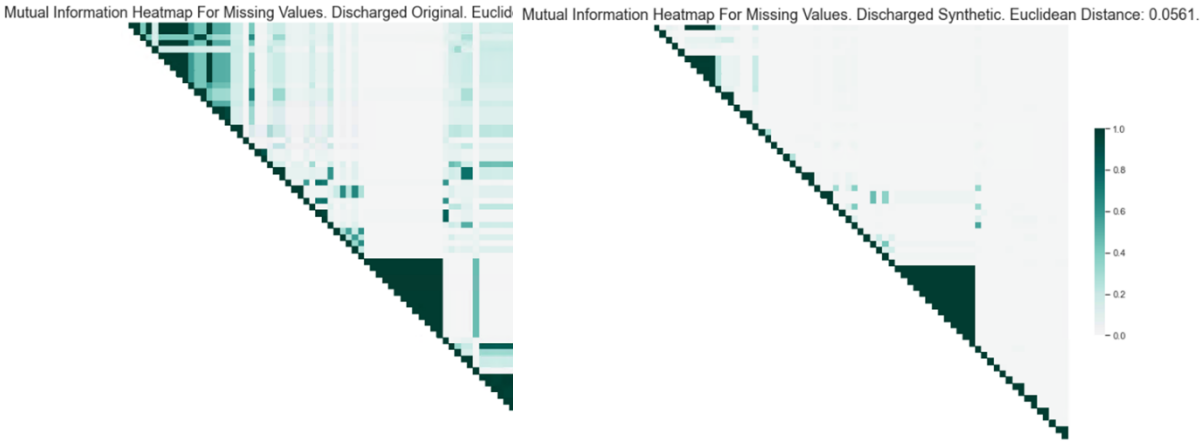


Euclidean distance: 0.1922

Interpretation: The fields with high positive correlations are preserved in the synthetic datasets. Some points lose the colour sharpness indicating the correlation of missing values between those 2 specific features is small. The overall distance is small indicating the missing values patterns are preserved between the datasets.

We see the Euclidean distance increasing from **0.176** (model 1) to **0.192** (model 2), indicating the missing value patters are preserved better by model 1, however, the difference is not significant.

Mutual Information



Euclidean distance: 0.0561

Interpretation: The fields with strong mutual information values are kept in the synthetic datasets. In the same time, the less strong associations between features are lost (right side columns).

The Euclidean distance increased from **0.0558** (model 1) to **0.0561** (model 2) indicating a lower association of missing values between features, however the difference is very small. We can say that the overall distance is small indicating the missing values patterns are preserved in similar fashion by model 1 and model 2.

Risk of Re-Identification

Risk of re-identification is a relative metric to the original dataset characteristics. It indicates the minimum and maximum risk of the individuals from the synthetic dataset to be re-identified.

The score is influenced by number of: (1) unique values in each column (2) number of identical values per individual with any point in the original dataset.

Statistic	Measurement
RIR min	0.279
RIR max	0.489
RIR mean	0.383
RIR scale	[0,2]

Interpretation: All datapoints of the synthetic dataset are evaluated with a risk between 0.28 and 0.49. A synthetic individual with a risk of 0 means that there is datapoint in the original dataset that has any identical values. All synthetic individuals have a risk in the lower 24% of the overall risk. Based on our experience, this risk of re-identification is low.

Appendix 2: Model 3

Model 3 is based on CTGAN, available as part of **sdv** library.

Comparison datasets:

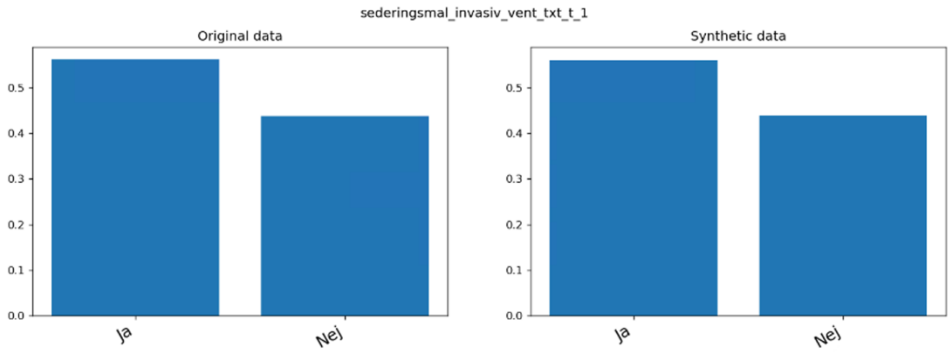
- Discharge_data_train.csv
- Discharged_synthetic_model3__1x_dataset.csv

Features similarity

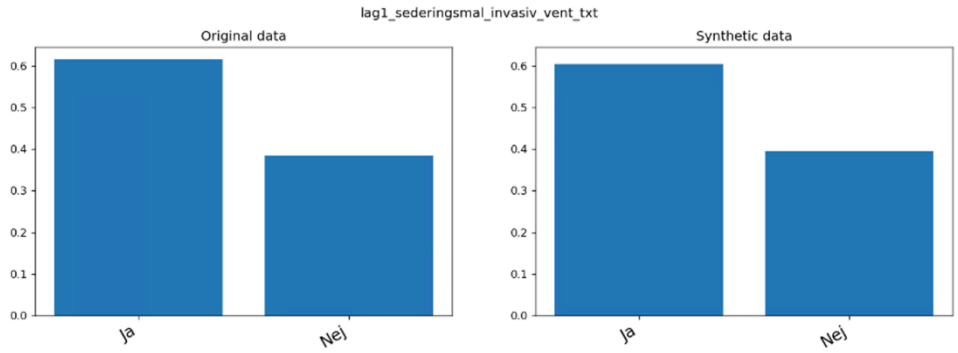
a. Individual distributions

Individual distributions visuals

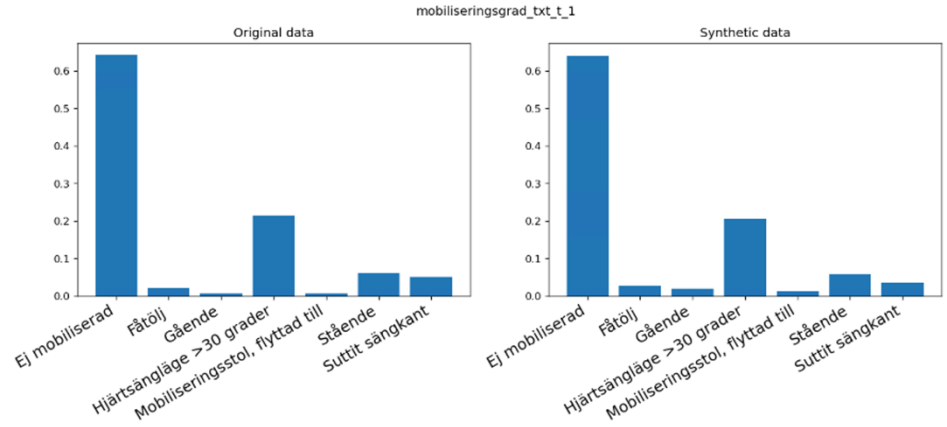
Examples of similar distributions



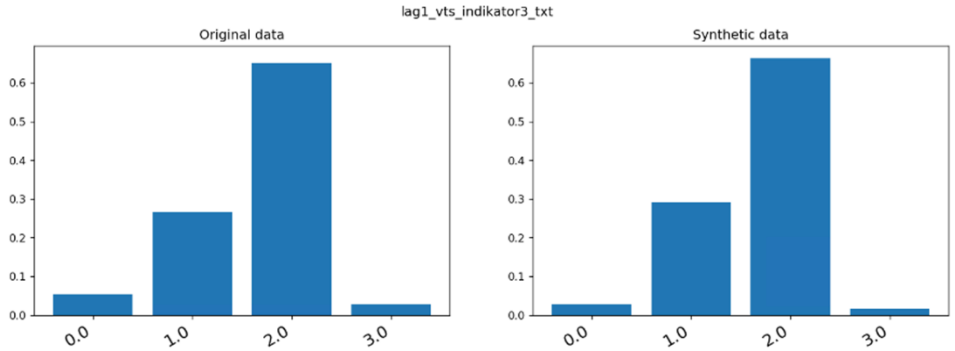
WD: 0.0006



WD: 0.0109



WD: 0.0118

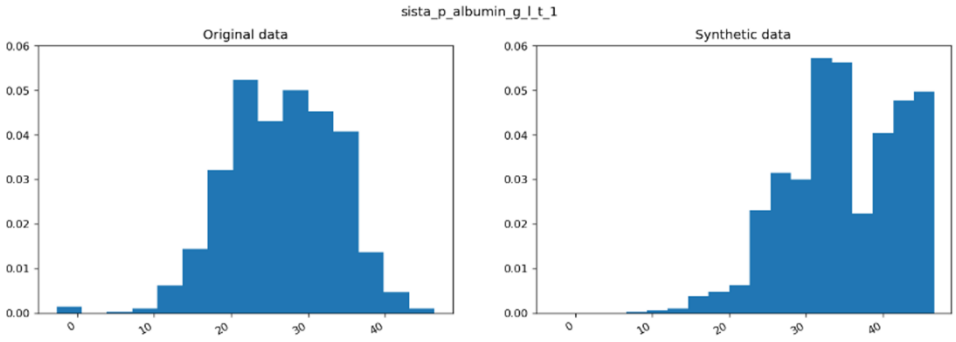


WD: 0.0128

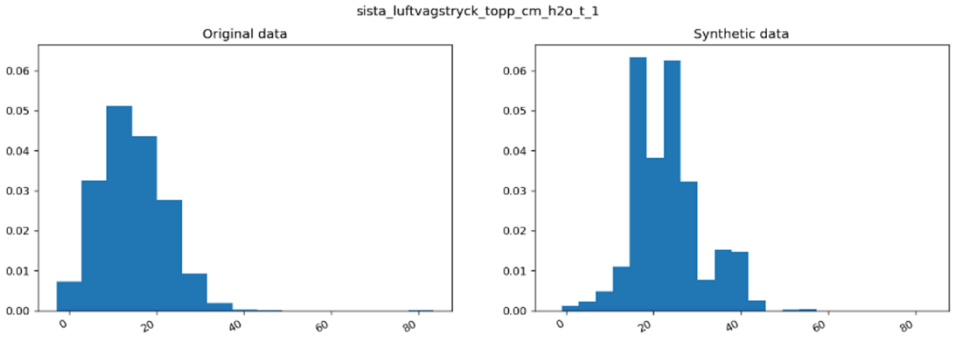
Interpretation: The most similar 5 features are categorical features with a WD close to 0 indicating a close replica to the original. These include *sederingsmal_invasiv_vent_txt_t_1, lag1_sederingsmal_invasiv_vent_txt, mobiliseringsgrad_txt_t_1, lag1_vts_indikator3_txt, vts_indikator9_cat_baseline.*

While model 1 and model 2 record the same 'best 5 features', model 3 produces different top features in terms of WD.

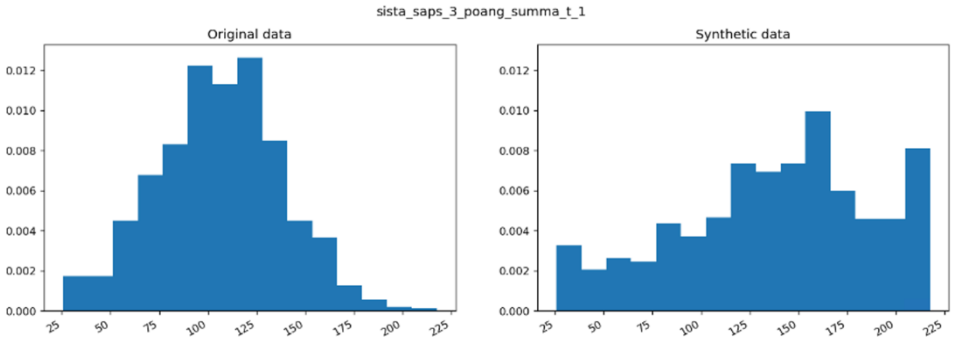
Examples of less similar distributions:



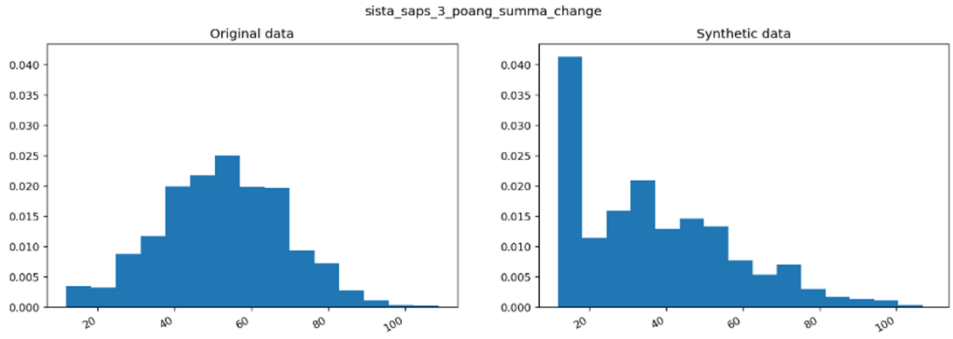
WD: 1.1371



WD: 1.058



WD: 1.037



WD: 1

Interpretation: The least similar 4 fields display WD below 1.13. Model 3 deems distributions with a higher WD in comparison to model 2.

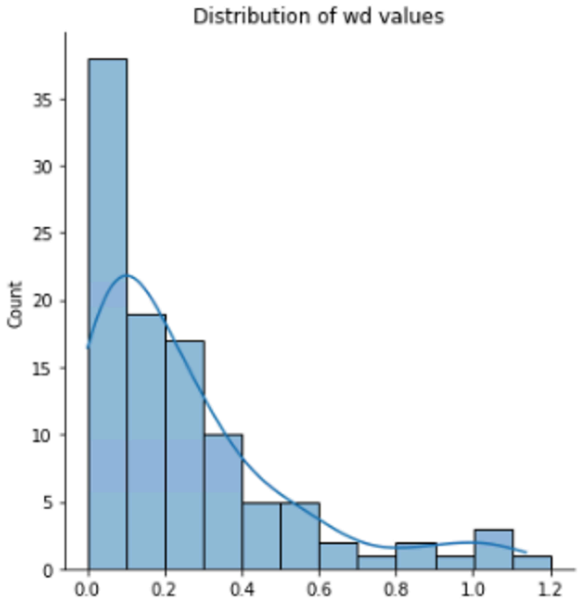
Model 3 reduces the similarity in distribution for the following features as compared to model 1: *sista_p_albumin_g_l_t_1, sista_luftvagstryck_topp_cm_h2o_t_1, sista_saps_3_poang_summa_t_1, sista_saps_3_poang_summa_change, sista_saps_3_poang_summa_baseline.*

In the same time, it improves the distributions for the following features: *Antal_mediciner_baseline, sista_saps_3_poang_flik_3_t_1, vts_medel_1, sista_p_albumin_g_l_baseline.*

The bottom worse distributions are not preserved between the models.

Overall WD distribution

```
General statistics of WD distribution:

WD Distribution Mean: 0.2514
WD Distribution Standard Deviation: 0.2623
```

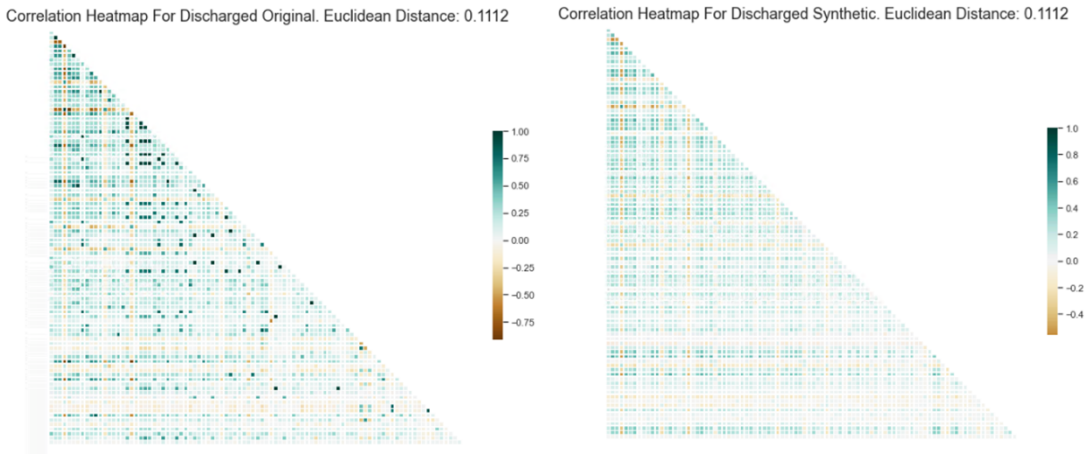


Interpretation:

As model 3 increased the WD of some of the features, the overall WD upper bound increased to 1.2. By assessing the mean and standard deviation above, we can conclude that model 3 has similar mean and standard deviation as model 1. However, the individual distribution quality shows better metrics for model 1.

b. Pairwise distributions

Pearson's correlation

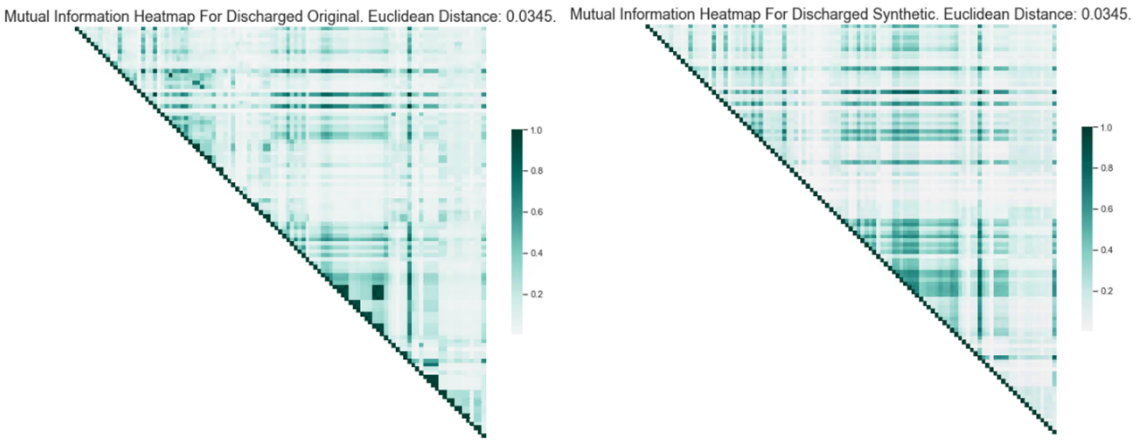


Euclidean distance: 0.1112

Interpretation:

Overall patterns are preserved. Model 3 shows an improvement in preserving some of the strong positive correlations leading to an improved Euclidean distance from **0.1187** (model 1) to **0.1112** (model 3). Overall, model 3 brings an improvement in correlations as compared to model 1.

Mutual Information



Euclidean distance: 0.0345

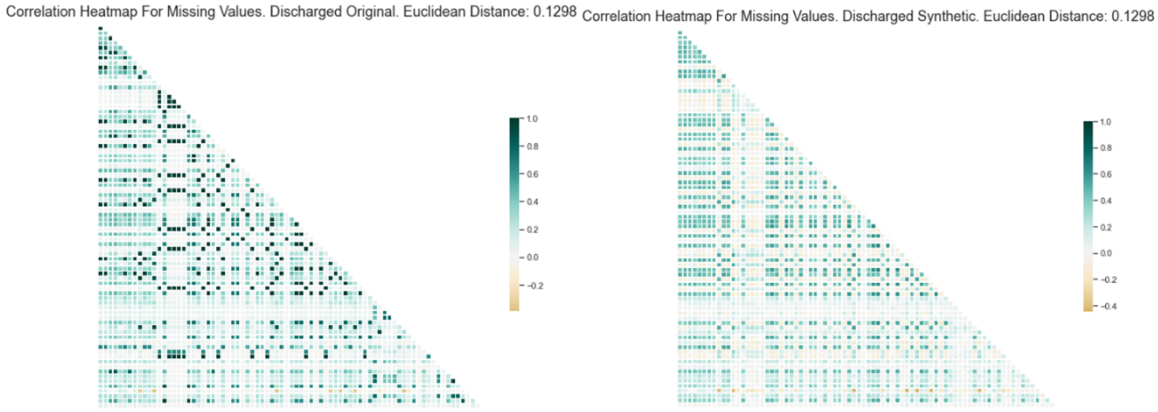
Interpretation: Pairwise distribution patterns are preserved between original and synthetic. Some of the strong mutual information values are preserved well. The overall distance is small indicating a high relatedness between fields.

Model 3 (**0.034**) outperforms model 2 (**0.046**). Model 3 (**0.034)** is similar in results as compared to model 1 (**0.033).**

Missing Values

a. Missing Values Matrices

Pearson's Correlation

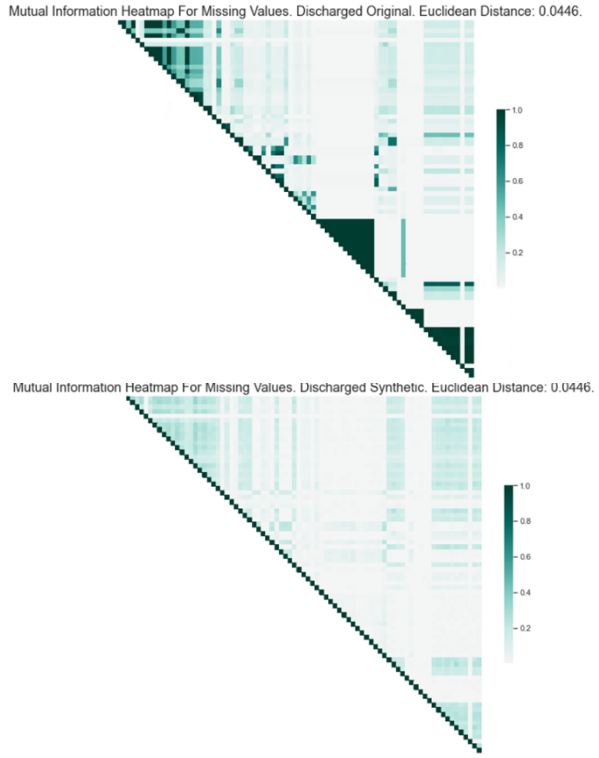


Euclidean distance: 0.1298

Interpretation: The fields with high positive correlations are preserved in the synthetic datasets. Some points lose the colour sharpness indicating the correlation of missing values between those 2 specific features is small. The overall distance is small indicating the missing values patterns are preserved between the datasets.

Model 3 outperforms the other 2 models in preserving the missing values patterns. We see the Euclidean distance decreased from **0.176** (model 1), **0.192** (model 2) to **0.129**(model 3), indicating the missing value patters are preserved better by model 3.

Mutual Information



Euclidean distance: 0.0446

Interpretation: The fields with strong mutual information values are not kept in the synthetic datasets. However, the less strong associations between features are preserved (right side columns).

The overall distance improved significantly. The Euclidean distance decreased from **0.0558** (model 1) and **0.0561** (model 2) to **0.0446** (model 3) indicating a better association of missing values between features. We can clearly see the missing values patterns in the right-side columns being preserved better than in model 1 and model 2.

Risk of Re-Identification

Risk of re-identification is a relative metric to the original dataset characteristics. It indicates the minimum and maximum risk of the individuals from the synthetic dataset to be re-identified.

The score is influenced by number of: (1) unique values in each column (2) number of identical values per individual with any point in the original dataset.

Statistic	Measurement
RIR min	0.17
RIR max	0.51
RIR mean	0.35
RIR scale	[0,2]
RIR execution time	2846s

Interpretation: All datapoints of the synthetic dataset are evaluated with a risk between 0.17 and 0.51. A synthetic individual with a risk of 0 means that there is datapoint in the original dataset that has any identical values. All synthetic individuals have a risk in the lower 25% of the overall risk. Based on our experience, this risk of re-identification is low.

Models' comparison

Model 1 (based on CTGAN networks) outperforms model 2 in better preserving missing values patterns. An explanation for this behaviour is that model 1 inputs fewer missing values than model 2. In contrast, model 2 has a higher rate of replacing missing values, thus the missing values pattern changes.

In addition, model 1 shows a small improvement in the Mutual Information score for the pairwise distributions, indicating that model 1 is able to capture some features relatedness better than model 2. However, the Pearson's correlation indicates the opposite.

Model 2 (based on Gaussian Copula) outperforms model 1 by producing very similar individual distributions and maintaining a better correlation between pairwise distributions, as recorded by Pearson's correlation score.

Model 3 (based on CTGAN) outperforms both model 1 and model 2 in terms of missing values preservation. Model 3 records better results on both Mutual Information and Pearson's Correlation metric).

The correlation metrics of model 3 show improved results in comparison with model 1. However, model 2 records higher levels of correlations.

In terms of individual distributions, model 3 has similar results as model 1. However, model 2 displays lower WD indicating better synthetisation of individual distributions.

During the improvement process of the models, it has been brought into attention that one of the features, namely *facit_vardlangd_kvar*, records large dissimilarities in comparison to the real dataset. **Model 3** was able to improve the statistics of *facit_vardlangd_kvar* feature and preserve the *facit_alvide2* ratio. One of the main advantages of **model 3** is the preserved ratio for *facit_alvide2==2* (150 datapoints in real vs 128 datapoints in synthetic dataset). This comes in opposition to **model 2** which disregards the proportion by producing only 5 datapoints satisfying *facit_alvide2==2* condition.

The table below summarises the similarity between the synthetic *facit_vardlangd_kvar* feature and the real one.

Statistics (facit_vardlangd_kvar)	Real	Synthetic
count	1689	1690
mean	56	54
std	107	149
min	0.38	0.38
25%	6.91	3.6
50%	13.2	11.85
75%	56.56	36.52
max	1673	1673
COUNT when facit_alviden == 2	150	128
MEAN when facit_alviden == 2	60	65
COUNT when facit_alviden == 1	1537	1549
MEAN when facit_alviden == 1	56	53