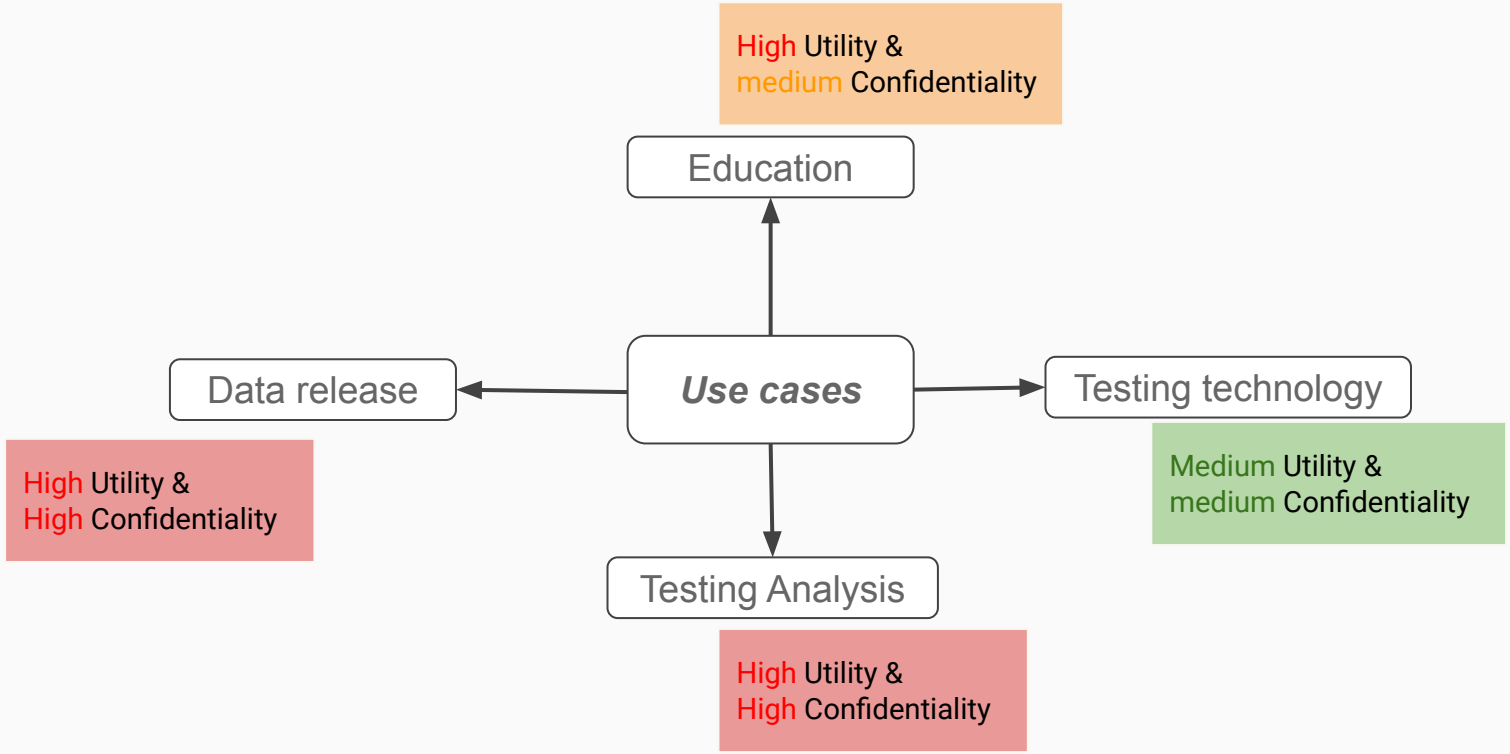
Statistics Netherlands Team

Satgpa Data set

Outline

- Understanding of the data sets (SATGPA, ACS)
- Partially synthetic data generation
  - Analytical Validity
  - Privacy evaluation
  - Utility evaluation
- Fully synthetic data generation
  - Analytical Validity
  - Privacy evaluation
  - Utility evaluation
- Judging the guideline

Synthetic Data Use Cases

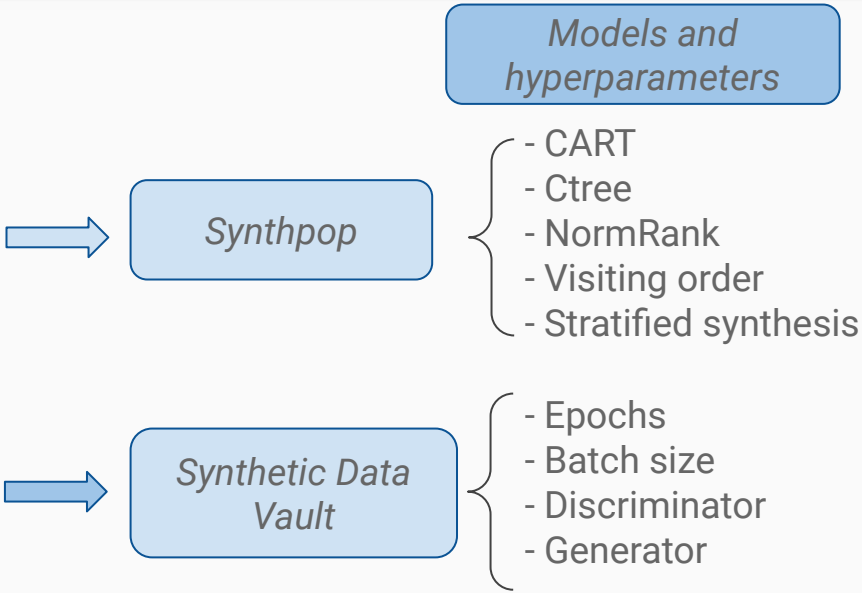


Data Synthesis Techniques

- Partially vs Fully synthetic data
- Sequential modeling
  - Fully conditional specification method (FCS)

- Deep Learning
  - Generative adversarial networks
  - Variational AutoEncoder
  - Gaussian Copulas

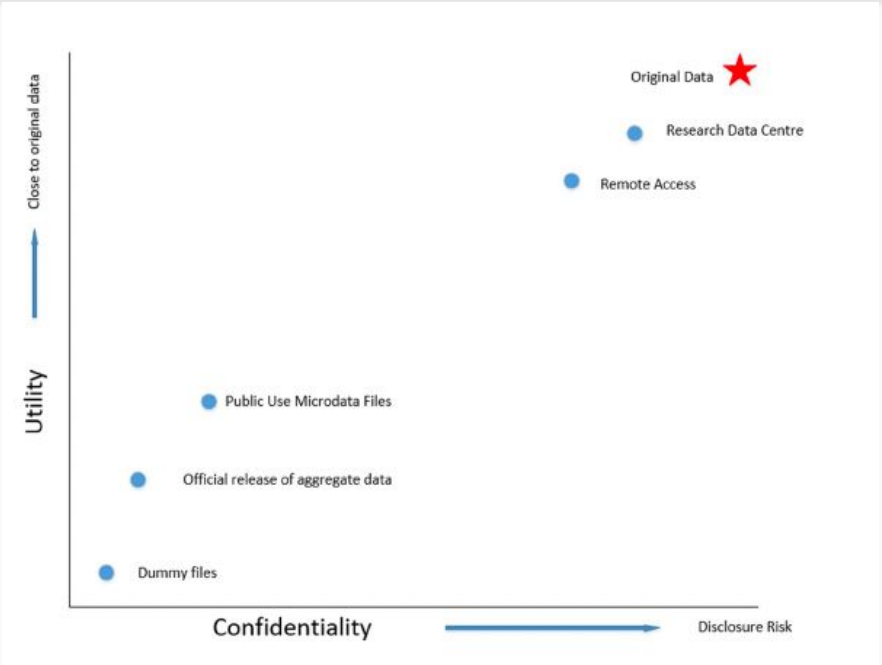
- Random data



Evaluation of synthetic data

How good is the generated data?

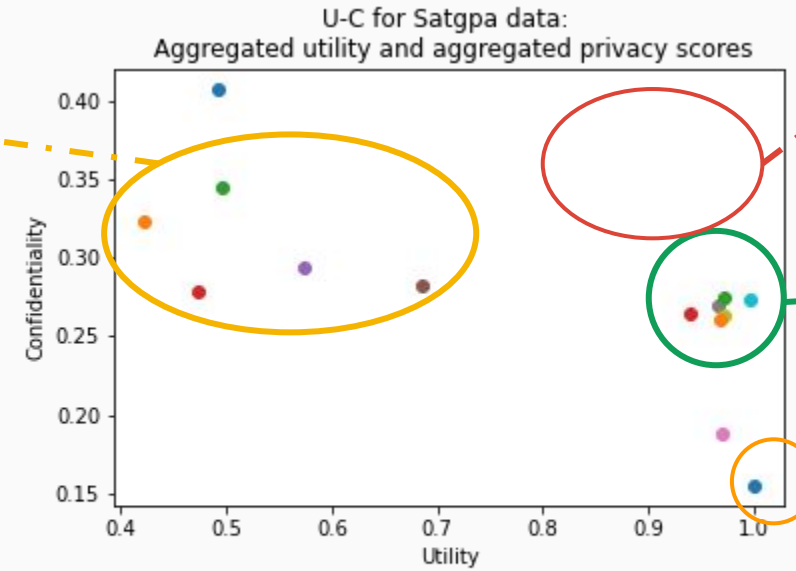
How much good is good?



***Credits**: Figure from chapter Introduction of the starter guide.*

Utility vs. Confidentiality on Satgpa real vs synthetic data sets

Testing Technology



Data Release ?

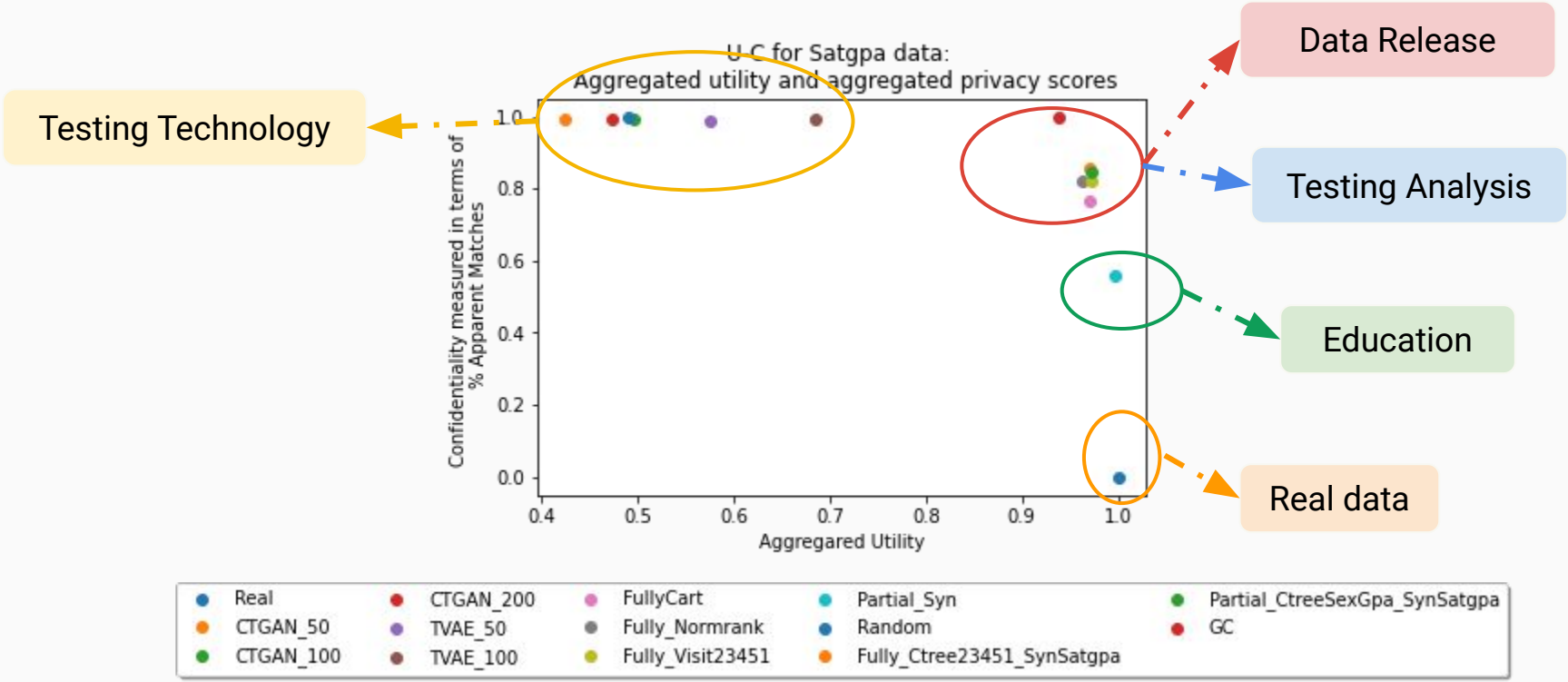
Testing Analysis ?

Education

Real data

Real
CTGAN_50
CTGAN_100
CTGAN_200
TVAE_50
TVAE_100
FullyCart
Fully_Normrank
Fully_Visit23451
Partial_Syn
Random
Fully_Ctree23451_SynSatgpa
Partial_CtreeSexGpa_SynSatgpa
GC

Utility vs. Confidentiality on Satgpa real vs synthetic data sets



Utility ~ Relative ranking of algorithms

	Real	CTGAN_50	CTGAN_100	CTGAN_200	TVAE_50	TVAE_100	FullyCart	Fully_Normrank	Fully_Visit23451
BinaryDecisionTreeClassifier	0.570 3	0.655 1	0.585 2	0.669 1	0.623 2	0.631 2	0.661 2	0.578 3	0.585 3
BinaryAdaBoostClassifier	0.603 2	0.582 2	0.623 1	0.632 2	0.613 3	0.624 3	0.641 3	0.650 2	0.602 2
BinaryLogisticRegression	0.664 1	0.566 3	0.502 3	0.301 3	0.689 1	0.677 1	0.667 1	0.651 1	0.651 1

Partial_Syn	Random	Fully_Ctree23451_SynSatgpa	Partial_CtreeSexGpa_SynSatgpa	GC
0.541 2	0.427 3	0.580 3	3 0.525	0.562 3
0.529 3	0.488 2	0.603 2	2 0.608	0.600 2
0.661 1	0.678 1	0.656 1	1 0.632	0.645 1

Utility Evaluation using SDV and Synthpop

- **Statistical Metrics**: These are metrics that compare the tables by running different statistical tests on them. Some of them work by comparing multiple columns at once, while other compare the different individual columns separately and later on return an aggregated result.
- **Likelihood Metrics**: These metrics attempt to fit a probabilistic model to the real data and later on evaluate the likelihood of the synthetic data on it.
- **Detection Metrics**: These metrics try to train a Machine Learning Classifier that learns to distinguish the real data from the synthetic data, and report a score of how successful this classifier is.
- **Machine Learning Efficacy Metrics**: These metrics train a Machine Learning model on your synthetic data and later on evaluate the model performance on the real data. Since these metrics need to evaluate the performance of a Machine Learning model on the dataset, they work only on datasets that represent a Machine Learning problem.

Source: https://sdv.dev/SDV/user_guides/evaluation/single_table_metrics.html

Privacy Evaluation

- Privacy metrics for re-identification attack aim to identify real users' records in the synthetic data
  - Apparent matches
  - Number of replicates
  - Distance to nearest neighbors
- Privacy metrics for *inference attack*: aim to infer sensitive information about real users from synthetic data
- Machine Learning efficacy:
  - $\mathcal{TRTS}$: Train Real, Test Synthetic
  - $\mathcal{TSTR}$: Train Synthetic, Test Real
  - $\mathcal{TRTR}$: Train Real, Test Real
  - $\mathcal{TSTS}$: Train Synthetic, Test Synthetic
  - $\mathcal{TSTS}$: Train Synthetic, Test Synthetic
  - $\mathcal{TMTM}$: Train and Test on Mixture of Real and Synthetic data.

Partially synthetic data

1. Partial synthesize for “sex” attribute using CART on Synthpop
2. Partial synthesize for “sex, hs_gpa, fy_gpa” using Ctrea on Synthpop

Satgpa~ Analytical Validity

Satgpa ~ syntheize of sex attribute only!

Interpretations:

-

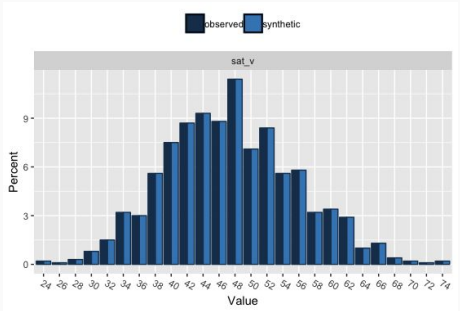
```
> summary(synthetic_dataset)
      sex            sat_v           sat_m          sat_sum         hs_gpa          fy_gpa
 Min.   :1.00   Min.   :24.0   Min.   :29.0   Min.   : 53   Min.   :1.8   Min.   :0.00
 1st Qu.:1.00   1st Qu.:43.0   1st Qu.:49.0   1st Qu.: 93   1st Qu.:2.8   1st Qu.:1.98
 Median :1.00   Median :49.0   Median :55.0   Median :103   Median :3.2   Median :2.46
 Mean   :1.48   Mean   :48.9   Mean   :54.4   Mean   :103   Mean   :3.2   Mean   :2.47
 3rd Qu.:2.00   3rd Qu.:54.0   3rd Qu.:60.0   3rd Qu.:113   3rd Qu.:3.7   3rd Qu.:3.02
 Max.   :2.00   Max.   :76.0   Max.   :77.0   Max.   :144   Max.   :4.5   Max.   :4.00
> summary(satgpa)
      sex            sat_v           sat_m          sat_sum         hs_gpa          fy_gpa
 Min.   :1.00   Min.   :24.0   Min.   :29.0   Min.   : 53   Min.   :1.8   Min.   :0.00
 1st Qu.:1.00   1st Qu.:43.0   1st Qu.:49.0   1st Qu.: 93   1st Qu.:2.8   1st Qu.:1.98
 Median :1.00   Median :49.0   Median :55.0   Median :103   Median :3.2   Median :2.46
 Mean   :1.48   Mean   :48.9   Mean   :54.4   Mean   :103   Mean   :3.2   Mean   :2.47
 3rd Qu.:2.00   3rd Qu.:54.0   3rd Qu.:60.0   3rd Qu.:113   3rd Qu.:3.7   3rd Qu.:3.02
 Max.   :2.00   Max.   :76.0   Max.   :77.0   Max.   :144   Max.   :4.5   Max.   :4.00
```

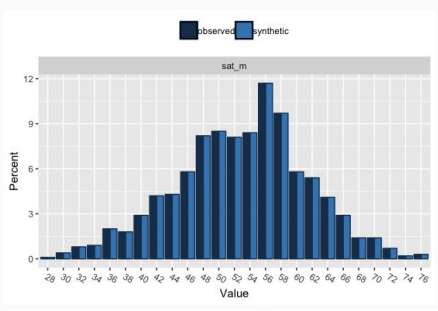
Satgpa ~ syntheize of sex attribute only!

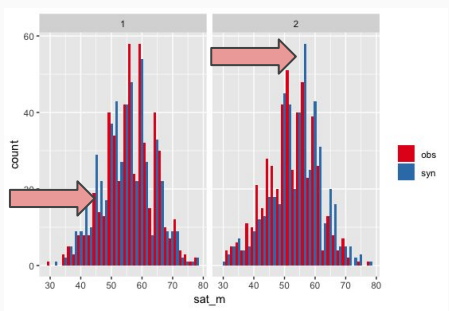
Interpretations:

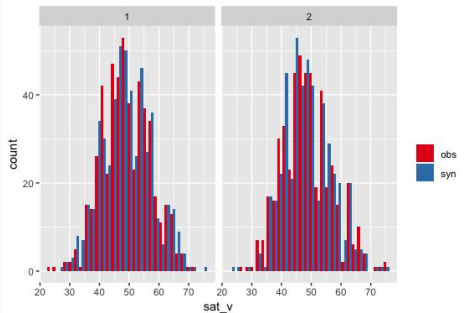
-

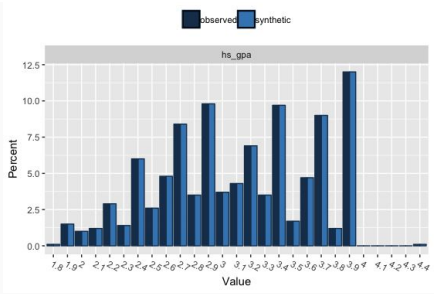
```
> utility.gen(synthetic_dataset, satgpa)
Running 50 permutations to get NULL utilities and printing every 10th.
synthesis 1   10 20 30 40 50
Utility score calculated by method: cart

Call:
utility.gen.data.frame(object = synthetic_dataset, data = satgpa)

Null utilities simulated from a permutation test with 50 replications.

Selected utility measures
   pMSE  S_pMSE
0.02333 0.80043
> utility.tables(synthetic_dataset, satgpa)

Two-way utility: S_pMSE value plotted for 15 pairs of variables.

Variable combinations with worst 5 utility scores (S_pMSE):
  1.sex:3.sat_m 1.sex:4.sat_sum  1.sex:6.fy_gpa  1.sex:5.hs_gpa   1.sex:2.sat_v
         6.6752          2.4931          2.0060          1.9719          0.8241

Medians and maxima of selected utility measures for all tables compared
       Medians  Maxima
pMSE         0  0.0038
S_pMSE       0  6.6752
df          24 24.0000
```

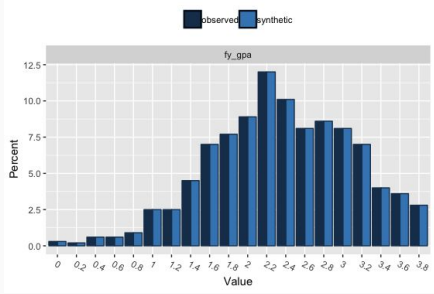

observed
synthetic
sat_v
Percent
Value

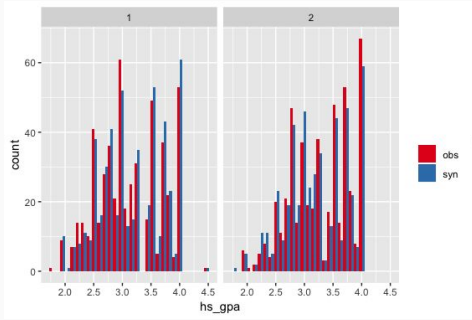

observed
synthetic
sat_m
Percent
Value

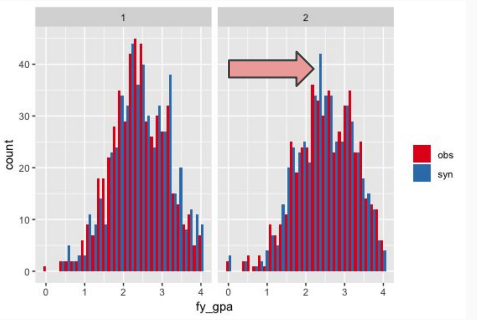

1
2
count
sat_m
obs
syn

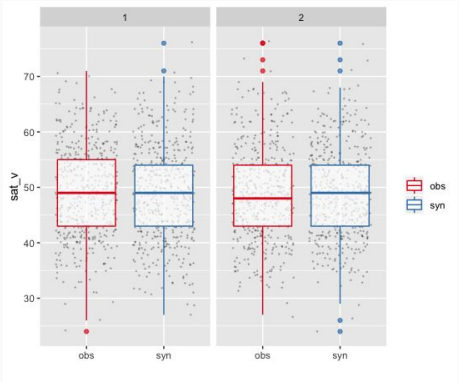

1
2
count
sat_v
obs
syn

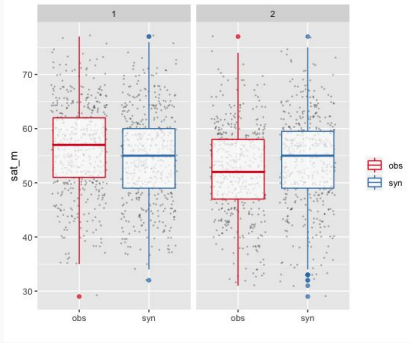

observed
synthetic
hs_gpa
Percent
Value

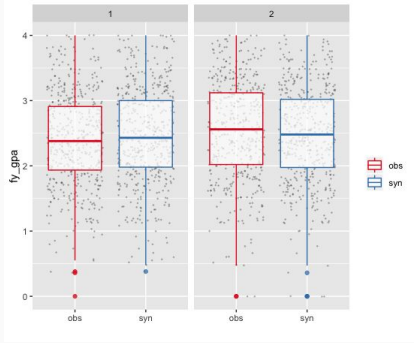

observed
synthetic
fy_gpa
Percent
Value

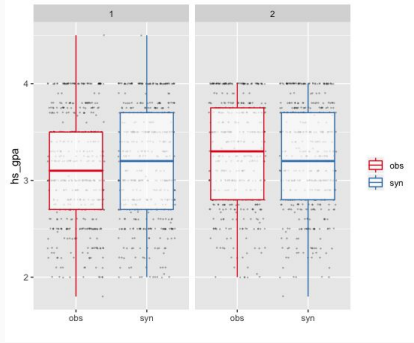

1
2
count
hs_gpa
obs
syn


1
2
count
fy_gpa
obs
syn


1
2
sat_v
obs
syn


1
2
sat_m
obs
syn

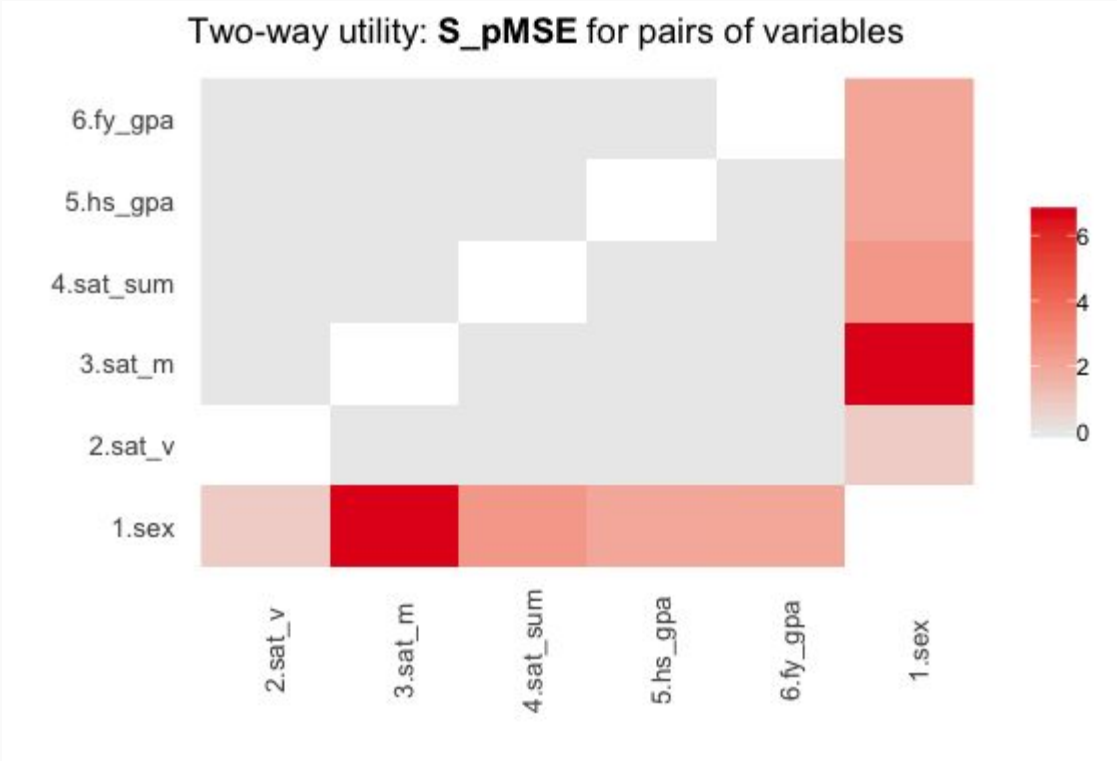

1
2
fy_gpa
obs
syn


1
2
hs_gpa
obs
syn

Satgpa ~ syntheize of sex attribute only!

Interpretations:

-



Satgpa~ Privacy Evaluation

- Apparent matching
- Duplicates
- Inference of sex

Privacy Evaluation

Replicated Uniques : Determines which unique units in the synthesised data set(s) replicates unique units in the original observed data set.

Interpretations:

- There are 510 replications!
- Percentage= 51 %

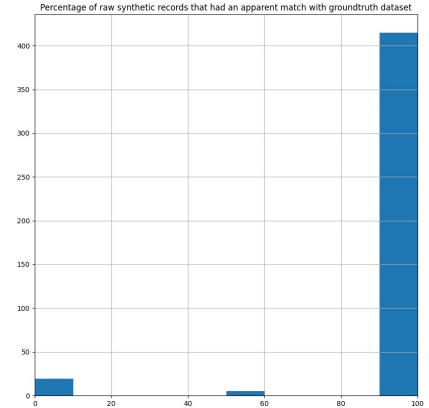
```
$no.uniques
[1] 1000

$no.replications
[1] 510

$per.replications
[1] 51
```

Apparent Match Distribution

Record similarity distribution between pairs of apparently matching records, using partially synthesis of satgpa data. There is no apparent unique matches between the real and synthetic data.

```
-q "sex, hs_gpa, fy_gpa"
 -x "sat_sum"
```

% of apparent matches= **43.9** %



Satgpa~ Utility Evaluation

- Machine Learning efficacy
- Relative Ranking of algorithms
- Likelihood / detective metrics

SDV metrics for different Methods

	Real	CTGAN_50	CTGAN_100	CTGAN_200	TVAE_50	TVAE_100	FullyCart	Fully_Normrank	Fully_Visit23451	Partial_Syn	Random	Fully_Ctree23451_SynSatgpa
LogisticDetection	1.000	0.196	0.296	0.301	0.354	0.502	1.000	0.991	1.000	1.000	0.353	1.000
KSTest	1.000	0.674	0.748	0.734	0.722	0.818	0.978	0.981	0.975	1.000	0.739	0.982
ContinuousKLDivergence	1.000	0.402	0.442	0.382	0.647	0.733	0.935	0.918	0.942	0.993	0.377	0.926
Aggregate Util	3.000	1.272	1.486	1.417	1.722	2.053	2.913	2.891	2.917	2.993	1.469	2.908
CategoricalGeneralizedCAP	0.200	0.493	0.489	0.492	0.478	0.489	0.293	0.485	0.462	0.477	0.485	0.473
CategoricalKNN	0.296	0.493	0.492	0.485	0.484	0.483	0.331	0.486	0.481	0.481	0.487	0.459
CategoricalSVM	0.233	0.482	0.486	0.487	0.483	0.486	0.302	0.435	0.425	0.487	0.498	0.423
Aggregate Cat	0.729	1.468	1.467	1.464	1.445	1.458	0.926	1.406	1.368	1.445	1.470	1.355
NumericalMLP	0.074	0.178	0.158	0.071	0.098	0.078	0.072	0.072	0.074	0.070	0.318	0.067
NumericalLR	0.066	0.150	0.210	0.071	0.089	0.070	0.066	0.066	0.066	0.066	0.321	0.066
NumericalSVR	0.065	0.182	0.229	0.069	0.125	0.093	0.065	0.069	0.067	0.065	0.322	0.068
Aggregate Num	0.204	0.509	0.596	0.210	0.312	0.241	0.203	0.207	0.206	0.201	0.960	0.201
Aggregated Privacy Metric	0.156	0.330	0.344	0.279	0.293	0.283	0.188	0.269	0.262	0.274	0.405	0.259
Aggregated Utility Metric	1.000	0.424	0.495	0.472	0.574	0.684	0.971	0.964	0.972	0.998	0.490	0.969
BinaryDecisionTreeClassifier	0.570	0.655	0.585	0.669	0.623	0.631	0.661	0.578	0.585	0.541	0.427	0.580
BinaryAdaBoostClassifier	0.603	0.582	0.623	0.632	0.613	0.624	0.641	0.650	0.602	0.529	0.488	0.603
BinaryLogisticRegression	0.664	0.566	0.502	0.301	0.689	0.677	0.667	0.651	0.651	0.661	0.678	0.656

Fully synthetic data

Full synthesize using CART

Satgpa ~ Full synthesize using CART

Interpretations:

-

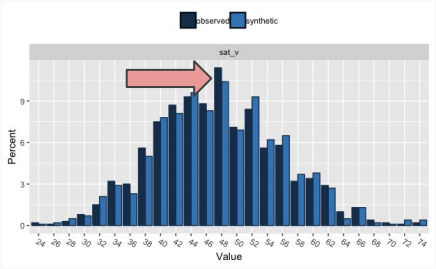
```
> summary(synthetic_dataset)
      sex            sat_v           sat_m          sat_sum         hs_gpa          fy_gpa
 Min.   :1.00   Min.   :26.0   Min.   :29.0   Min.   : 60   Min.   :1.80   Min.   :0.00
 1st Qu.:1.00   1st Qu.:43.0   1st Qu.:49.0   1st Qu.: 93   1st Qu.:2.80   1st Qu.:2.00
 Median :1.00   Median :49.0   Median :55.0   Median :104   Median :3.20   Median :2.52
 Mean   :1.49   Mean   :49.2   Mean   :54.5   Mean   :104   Mean   :3.21   Mean   :2.51
 3rd Qu.:2.00   3rd Qu.:55.0   3rd Qu.:60.0   3rd Qu.:113   3rd Qu.:3.70   3rd Qu.:3.07
 Max.   :2.00   Max.   :76.0   Max.   :77.0   Max.   :144   Max.   :4.50   Max.   :4.00
> summary(satgpa)
      sex            sat_v           sat_m          sat_sum         hs_gpa          fy_gpa
 Min.   :1.00   Min.   :24.0   Min.   :29.0   Min.   : 53   Min.   :1.8   Min.   :0.00
 1st Qu.:1.00   1st Qu.:43.0   1st Qu.:49.0   1st Qu.: 93   1st Qu.:2.8   1st Qu.:1.98
 Median :1.00   Median :49.0   Median :55.0   Median :103   Median :3.2   Median :2.46
 Mean   :1.48   Mean   :48.9   Mean   :54.4   Mean   :103   Mean   :3.2   Mean   :2.47
 3rd Qu.:2.00   3rd Qu.:54.0   3rd Qu.:60.0   3rd Qu.:113   3rd Qu.:3.7   3rd Qu.:3.02
 Max.   :2.00   Max.   :76.0   Max.   :77.0   Max.   :144   Max.   :4.5   Max.   :4.00
```

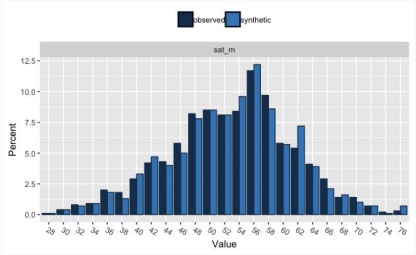
Satgpa ~ Full synthesize using CART

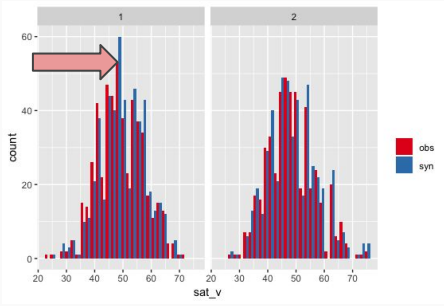
Interpretations:

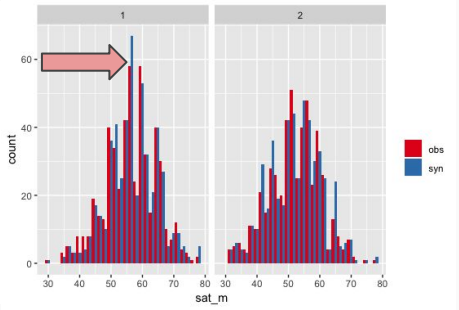
-

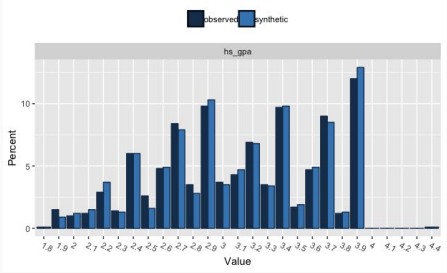
```
> utility.gen(synthetic_dataset, satgpa)
Running 50 permutations to get NULL utilities and printing every 10th.
synthesis 1   10 20 30 40 50
Utility score calculated by method: cart

Call:
utility.gen.data.frame(object = synthetic_dataset, data = satgpa)

Null utilities simulated from a permutation test with 50 replications.

Selected utility measures
   pMSE  S_pMSE
0.05523 1.64222
> utility.tables(synthetic_dataset, satgpa)

Two-way utility: S_pMSE value plotted for 15 pairs of variables.

Variable combinations with worst 5 utility scores (S_pMSE):
3.sat_m:4.sat_sum  2.sat_v:6.fy_gpa     1.sex:2.sat_v 2.sat_v:4.sat_sum  3.sat_m:6.fy_gpa
            1.601             1.590            1.507             1.433             1.279

Medians and maxima of selected utility measures for all tables compared
       Medians  Maxima
pMSE    0.0011  0.0024
S_pMSE  1.2148  1.6011
df     24.0000 24.0000

For more details of all scores use print.tabs = TRUE.
```

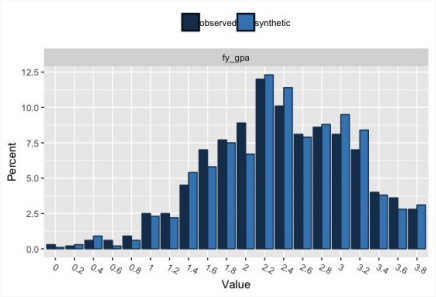

observed
synthetic
sat_v
Percent
Value

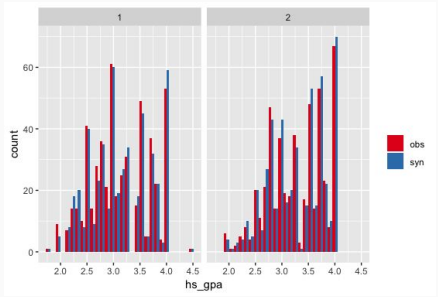

observed
synthetic
sat_m
Percent
Value

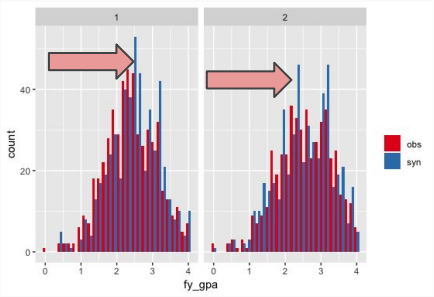

1
2
count
sat_v
obs
syn

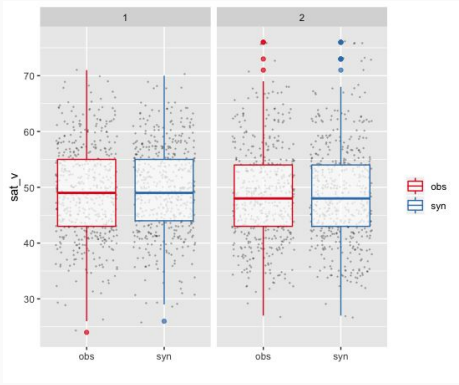

1
2
count
sat_m
obs
syn

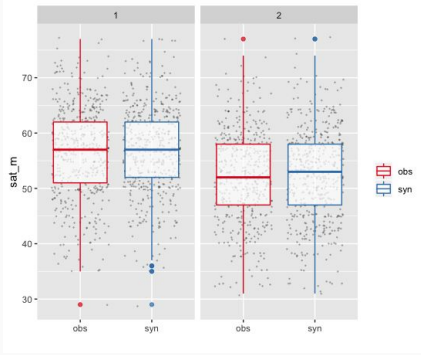

observed
synthetic
hs_gpa
Percent
Value

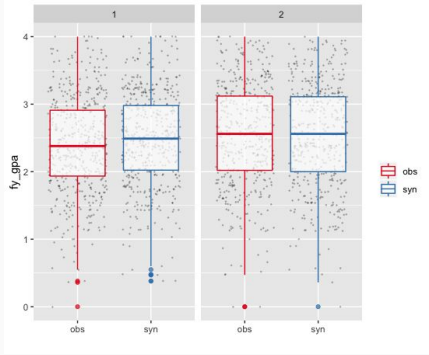

observed
synthetic
fy_gpa
Percent
Value

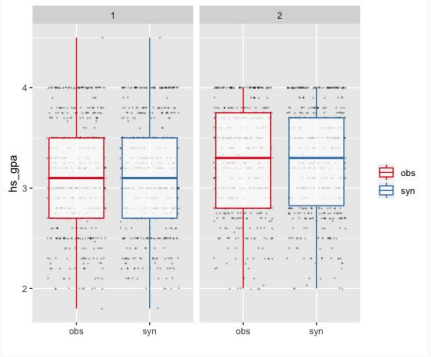

1
2
count
hs_gpa
obs
syn


1
2
count
fy_gpa
obs
syn


1
2
sat_v
obs
syn


1
2
sat_m
obs
syn

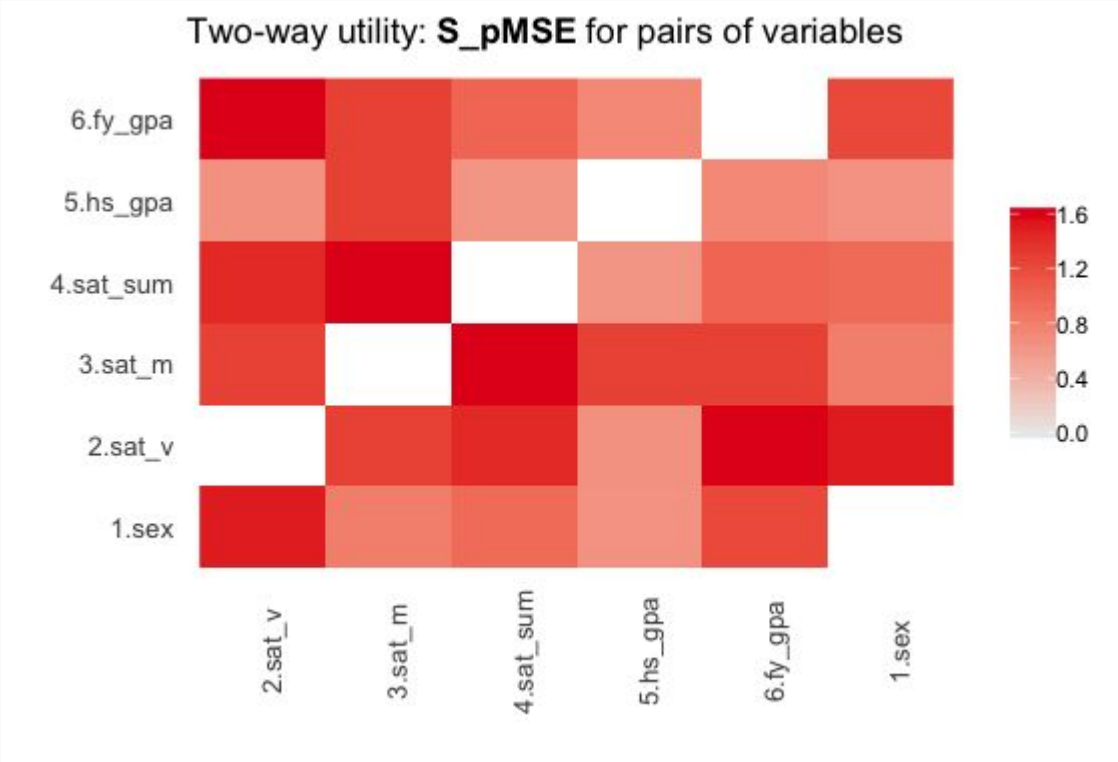

1
2
fy_gpa
obs
syn


1
2
hs_gpa
obs
syn

Satgpa ~ Full synthesize using CART

Interpretations:

-



Privacy Evaluation

Replicated Uniques : Determines which unique units in the synthesised data set(s) replicates unique units in the original observed data set.

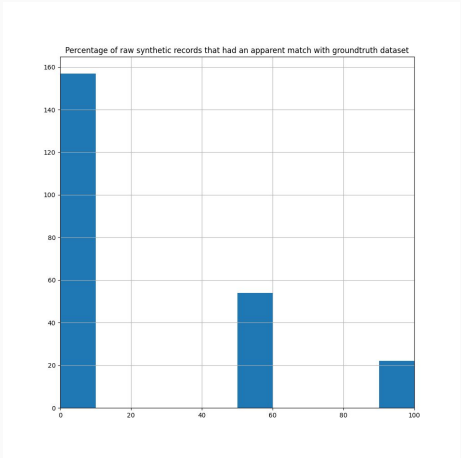
Interpretations:

- There are 34 replications!
- Percentage= 3.4 %

```
$no.uniques
[1] 1000

$no.replications
[1] 34

$per.replications
[1] 3.4
```

Apparent Match Distribution

Record similarity distribution between pairs of apparently matching records, using partially synthesis of satgpa data. There is no apparent unique matches between the real and synthetic data.



Full Synthesize using NormRank

Satgpa ~ synthesize All using Normrank

Interpretations:

-

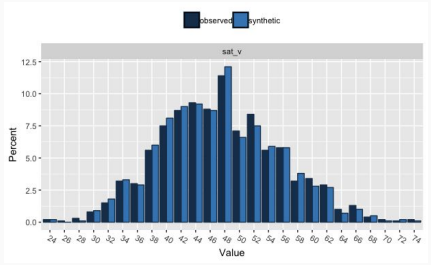
```
> summary(synthetic_dataset)
      sex            sat_v           sat_m          sat_sum         hs_gpa          fy_gpa
 Min.   :1.00   Min.   :26.0   Min.   :29.0   Min.   : 57   Min.   :2.00   Min.   :0.00
 1st Qu.:1.00   1st Qu.:43.0   1st Qu.:49.0   1st Qu.: 93   1st Qu.:2.75   1st Qu.:1.96
 Median :1.00   Median :48.0   Median :55.0   Median :103   Median :3.20   Median :2.42
 Mean   :1.49   Mean   :48.7   Mean   :54.2   Mean   :103   Mean   :3.19   Mean   :2.46
 3rd Qu.:2.00   3rd Qu.:54.0   3rd Qu.:60.0   3rd Qu.:112   3rd Qu.:3.70   3rd Qu.:3.00
 Max.   :2.00   Max.   :76.0   Max.   :77.0   Max.   :153   Max.   :4.50   Max.   :4.00
> summary(satgpa)
      sex            sat_v           sat_m          sat_sum         hs_gpa          fy_gpa
 Min.   :1.00   Min.   :24.0   Min.   :29.0   Min.   : 53   Min.   :1.8   Min.   :0.00
 1st Qu.:1.00   1st Qu.:43.0   1st Qu.:49.0   1st Qu.: 93   1st Qu.:2.8   1st Qu.:1.98
 Median :1.00   Median :49.0   Median :55.0   Median :103   Median :3.2   Median :2.46
 Mean   :1.48   Mean   :48.9   Mean   :54.4   Mean   :103   Mean   :3.2   Mean   :2.47
 3rd Qu.:2.00   3rd Qu.:54.0   3rd Qu.:60.0   3rd Qu.:113   3rd Qu.:3.7   3rd Qu.:3.02
 Max.   :2.00   Max.   :76.0   Max.   :77.0   Max.   :144   Max.   :4.5   Max.   :4.00
```

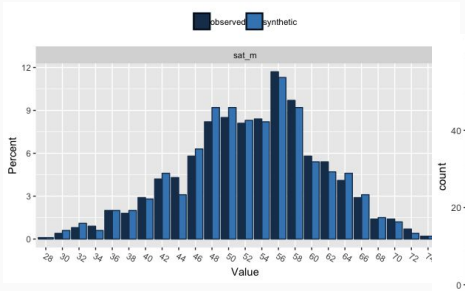
Satgpa ~ synthesize All using Normrank

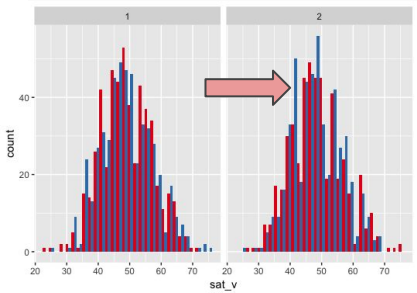
Interpretations:

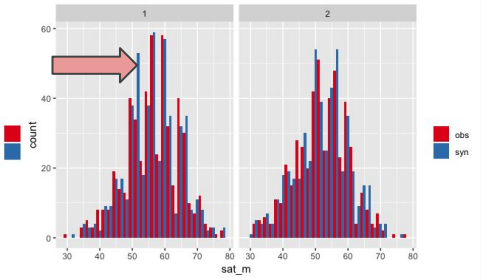
-

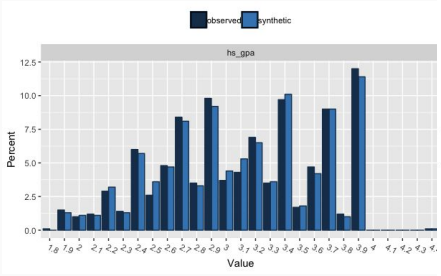
```
> utility.gen(synthetic_dataset, satgpa)
Running 50 permutations to get NULL utilities and printing every 10th.
synthesis 1   10 20 30 40 50
Utility score calculated by method: cart

Call:
utility.gen.data.frame(object = synthetic_dataset, data = satgpa)

Null utilities simulated from a permutation test with 50 replications.

Selected utility measures
  pMSE S_pMSE
0.0677 1.8706
> utility.tables(synthetic_dataset, satgpa)

Two-way utility: S_pMSE value plotted for 15 pairs of variables.

Variable combinations with worst 5 utility scores (S_pMSE):
   1.sex:4.sat_sum 4.sat_sum:5.hs_gpa  2.sat_v:4.sat_sum   2.sat_v:6.fy_gpa    2.sat_v:3.sat_m
             1.943              1.747             1.676             1.519              1.515

Medians and maxima of selected utility measures for all tables compared
       Medians  Maxima
pMSE    0.0014  0.0026
S_pMSE  1.0014  1.9432
df     24.0000 24.0000
```

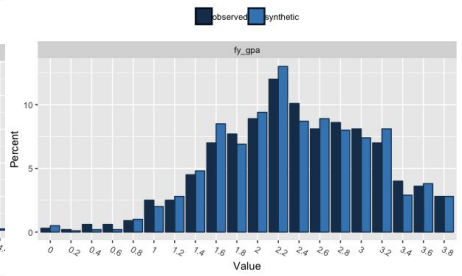

observed
synthetic
sat_v
Percent
Value

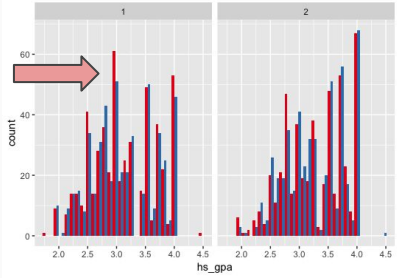

observed
synthetic
sat_m
Percent
Value

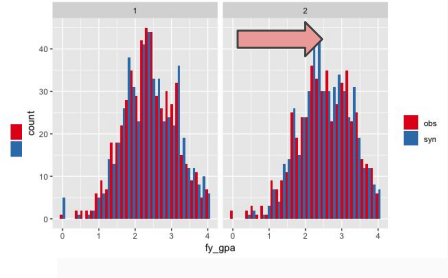

1
2
count
sat_v

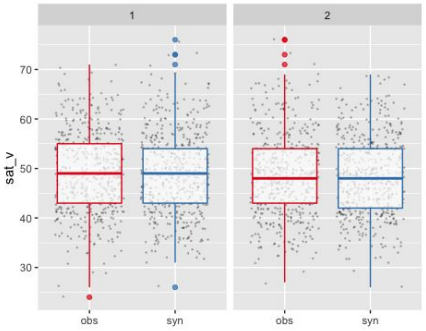

1
2
count
sat_m
obs
syn

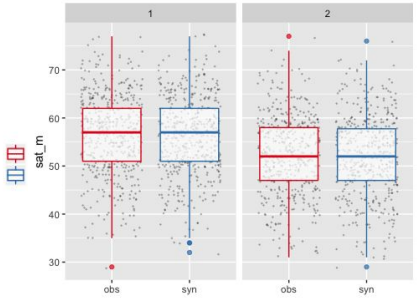

observed
synthetic
hs_gpa
Percent
Value

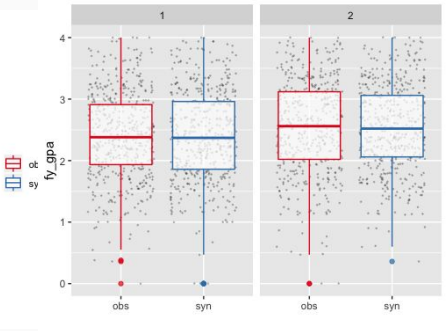

observed
synthetic
fy_gpa
Percent
Value

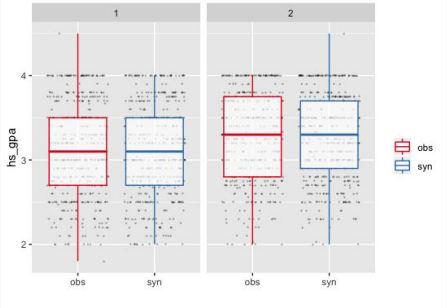

1
2
count
hs_gpa


1
2
count
fy_gpa
obs
syn


1
2
sat_v
obs
syn


1
2
sat_m
obs
syn

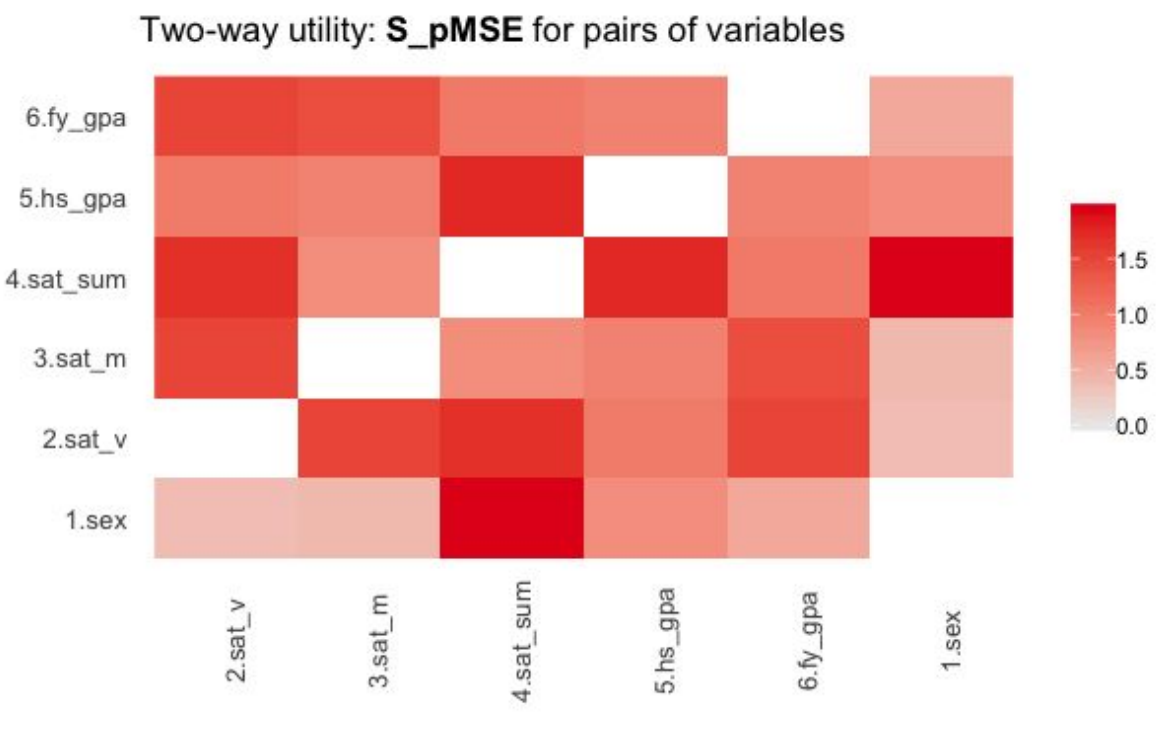

1
2
fy_gpa
obs
syn


1
2
hs_gpa
obs
syn

Satgpa ~ Full synthesize using Normrank

Interpretations:

-



Privacy Evaluation

Replicated Uniques : Determines which unique units in the synthesised data set(s) replicates unique units in the original observed data set.

Interpretations:

- There are 0 replications!
- Percentage= 0.0 %

```
$no.uniques
[1] 1000

$no.replications
[1] 0

$per.replications
[1] 0
```

Apparent Match Distribution
Record similarity distribution between pairs of apparently matching records, using partially synthesis of satgpa data. There is no apparent unique matches between the real and synthetic data.

CTGAN

Satgpa ~ Full synthesize using CTGAN

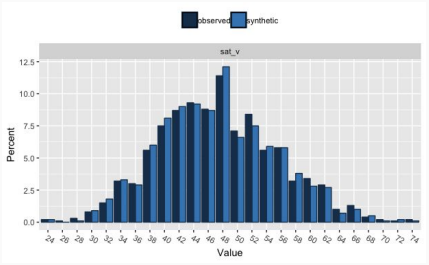
```
> summary(synthetic_dataset)
       X              sex             sat_v           sat_m          hs_gpa          fy_gpa          sat_sum     
 Min.   :  0   Min.   :1.00   Min.   :25.0   Min.   :27.0   Min.   :1.51   Min.   :0.215   Min.   : 59  
 1st Qu.:250   1st Qu.:1.00   1st Qu.:42.0   1st Qu.:47.0   1st Qu.:2.46   1st Qu.:2.083   1st Qu.: 92  
 Median :500   Median :1.00   Median :48.0   Median :55.0   Median :2.92   Median :2.665   Median :103  
 Mean   :500   Mean   :1.47   Mean   :48.5   Mean   :54.5   Mean   :2.96   Mean   :2.648   Mean   :103  
 3rd Qu.:749   3rd Qu.:2.00   3rd Qu.:55.0   3rd Qu.:63.0   3rd Qu.:3.54   3rd Qu.:3.331   3rd Qu.:113  
 Max.   :999   Max.   :2.00   Max.   :73.0   Max.   :80.0   Max.   :4.21   Max.   :4.225   Max.   :145  
> summary(satgpa)
      sex            sat_v           sat_m          sat_sum         hs_gpa          fy_gpa      
 Min.   :1.00   Min.   :24.0   Min.   :29.0   Min.   : 53   Min.   :1.8   Min.   :0.00  
 1st Qu.:1.00   1st Qu.:43.0   1st Qu.:49.0   1st Qu.: 93   1st Qu.:2.8   1st Qu.:1.98  
 Median :1.00   Median :49.0   Median :55.0   Median :103   Median :3.2   Median :2.46  
 Mean   :1.48   Mean   :48.9   Mean   :54.4   Mean   :103   Mean   :3.2   Mean   :2.47  
 3rd Qu.:2.00   3rd Qu.:54.0   3rd Qu.:60.0   3rd Qu.:113   3rd Qu.:3.7   3rd Qu.:3.02  
 Max.   :2.00   Max.   :76.0   Max.   :77.0   Max.   :144   Max.   :4.5   Max.   :4.00  
```

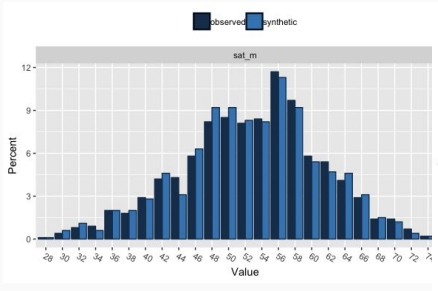
Satgpa ~ Full synthesize using CTGAN

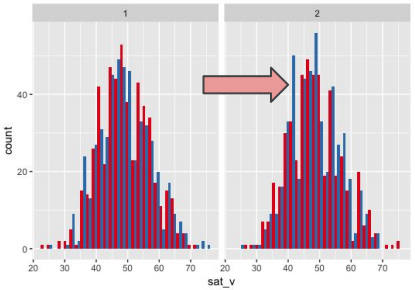
Interpretations:

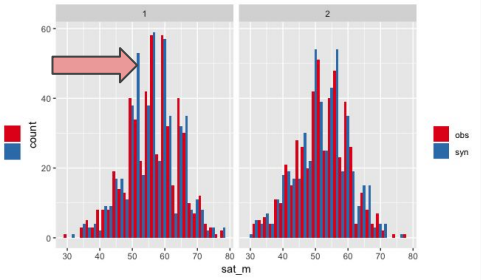
-

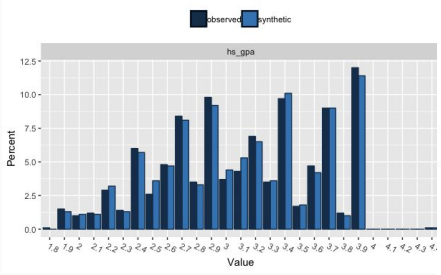
```
> utility.gen(synthetic_dataset, satgpa)
Running 50 permutations to get NULL utilities and printing every 10th.
synthesis 1   10 20 30 40 50
Utility score calculated by method: cart

Call:
utility.gen.data.frame(object = synthetic_dataset, data = satgpa)

Null utilities simulated from a permutation test with 50 replications.

Selected utility measures
  pMSE S_pMSE
0.2195 5.4668
> utility.tables(synthetic_dataset, satgpa)

Two-way utility: S_pMSE value plotted for 15 pairs of variables.

Variable combinations with worst 5 utility scores (S_pMSE):
    1.sex:4.hs_gpa  4.hs_gpa:5.fy_gpa   3.sat_m:4.hs_gpa 4.hs_gpa:6.sat_sum   2.sat_v:4.hs_gpa
             30.57              24.02              17.20              13.92              13.38

Medians and maxima of selected utility measures for all tables compared
       Medians Maxima
pMSE    0.0122  0.036
S_pMSE  9.4784 30.570
df     24.0000 24.000

For more details of all scores use print.tabs = TRUE.
```

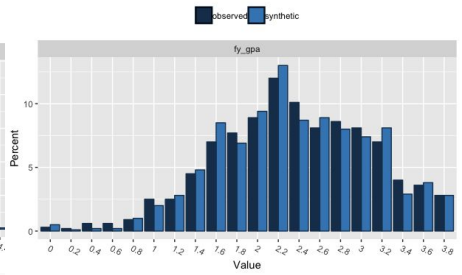

observed
synthetic
sat_v
Percent
Value

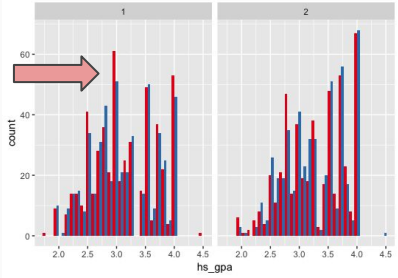

observed
synthetic
sat_m
Percent
Value

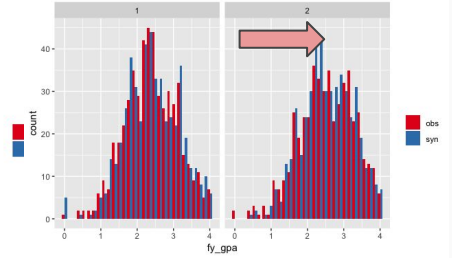

1
2
count
sat_v

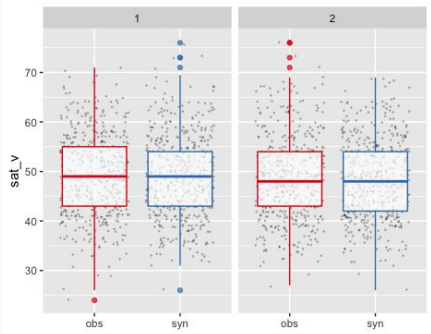

1
2
count
sat_m
obs
syn

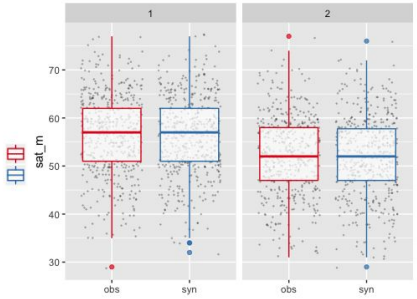

observed
synthetic
hs_gpa
Percent
Value

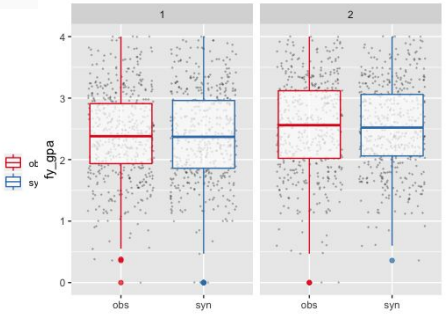

observed
synthetic
fy_gpa
Percent
Value

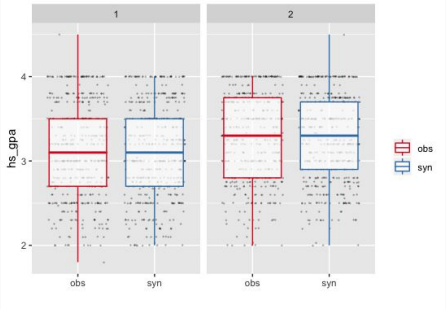

1
2
count
hs_gpa


1
2
count
fy_gpa
obs
syn


1
2
sat_v
obs
syn


1
2
sat_m
obs
syn

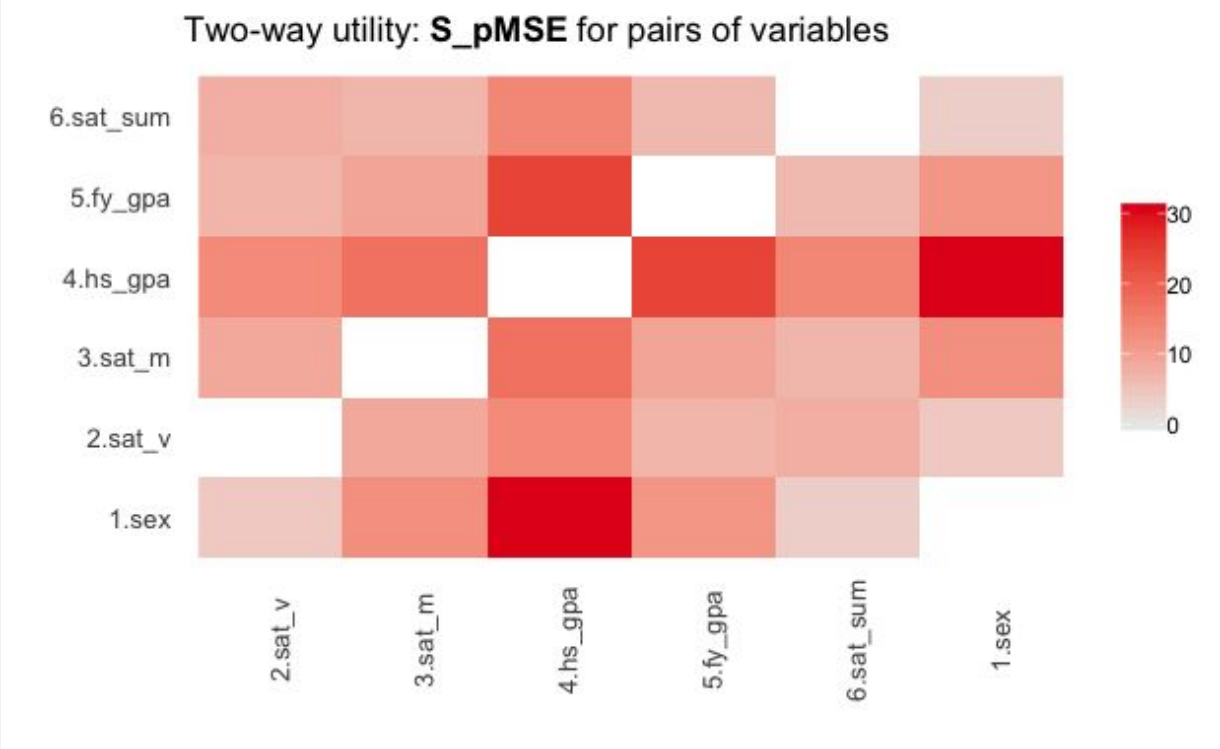

1
2
fy_gpa
obs
syn


1
2
hs_gpa
obs
syn

Satgpa ~ Full synthesize using CTGAN

Interpretations:

-



Privacy Evaluation

Replicated Uniques : Determines which unique units in the synthesised data set(s) replicates unique units in the original observed data set.

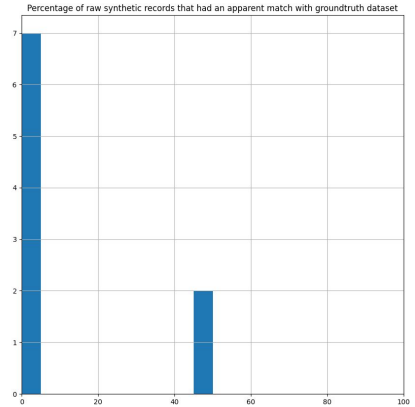
Interpretations:

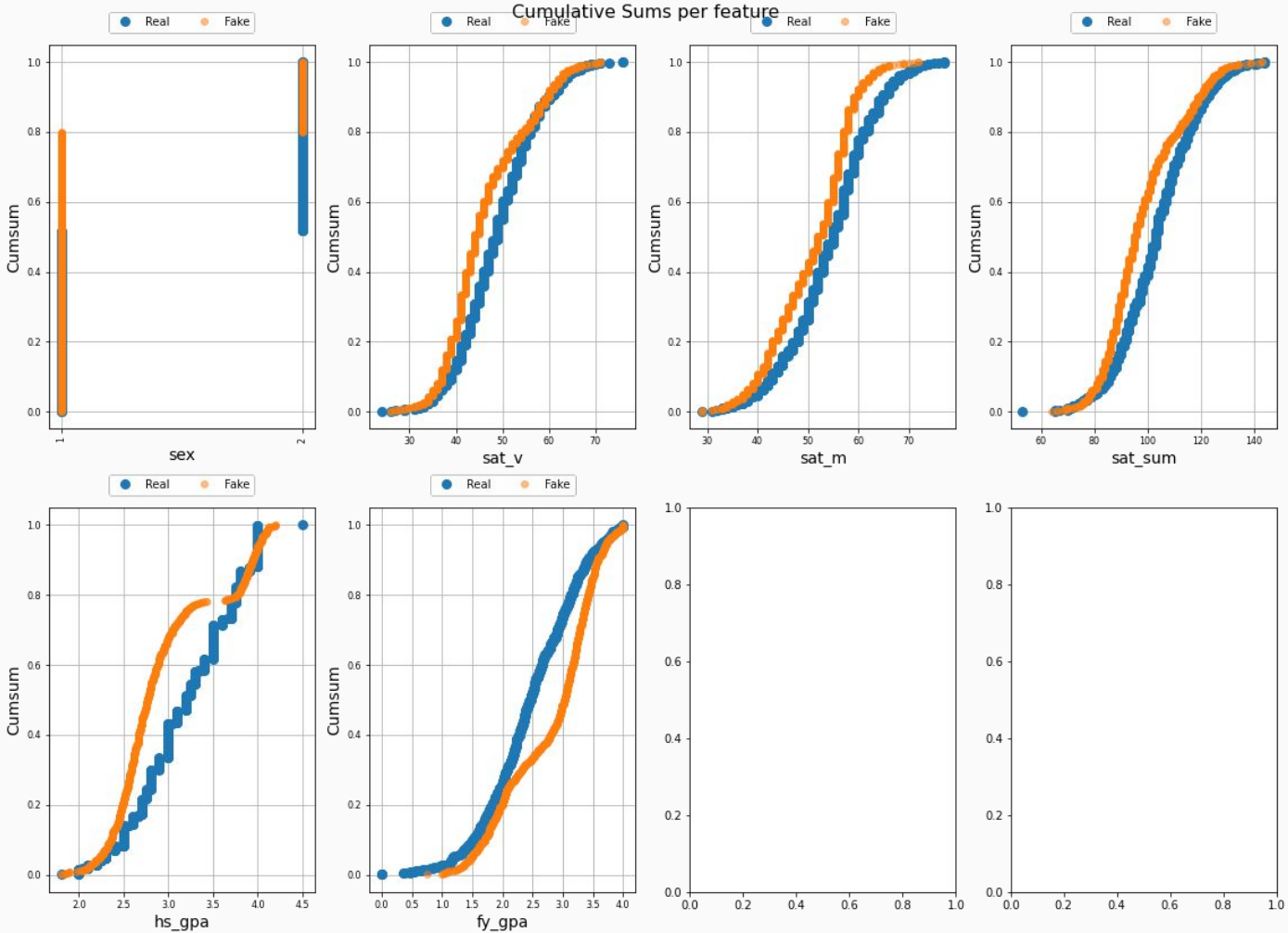
- There are 0 replications!
- Percentage= 0.0 %

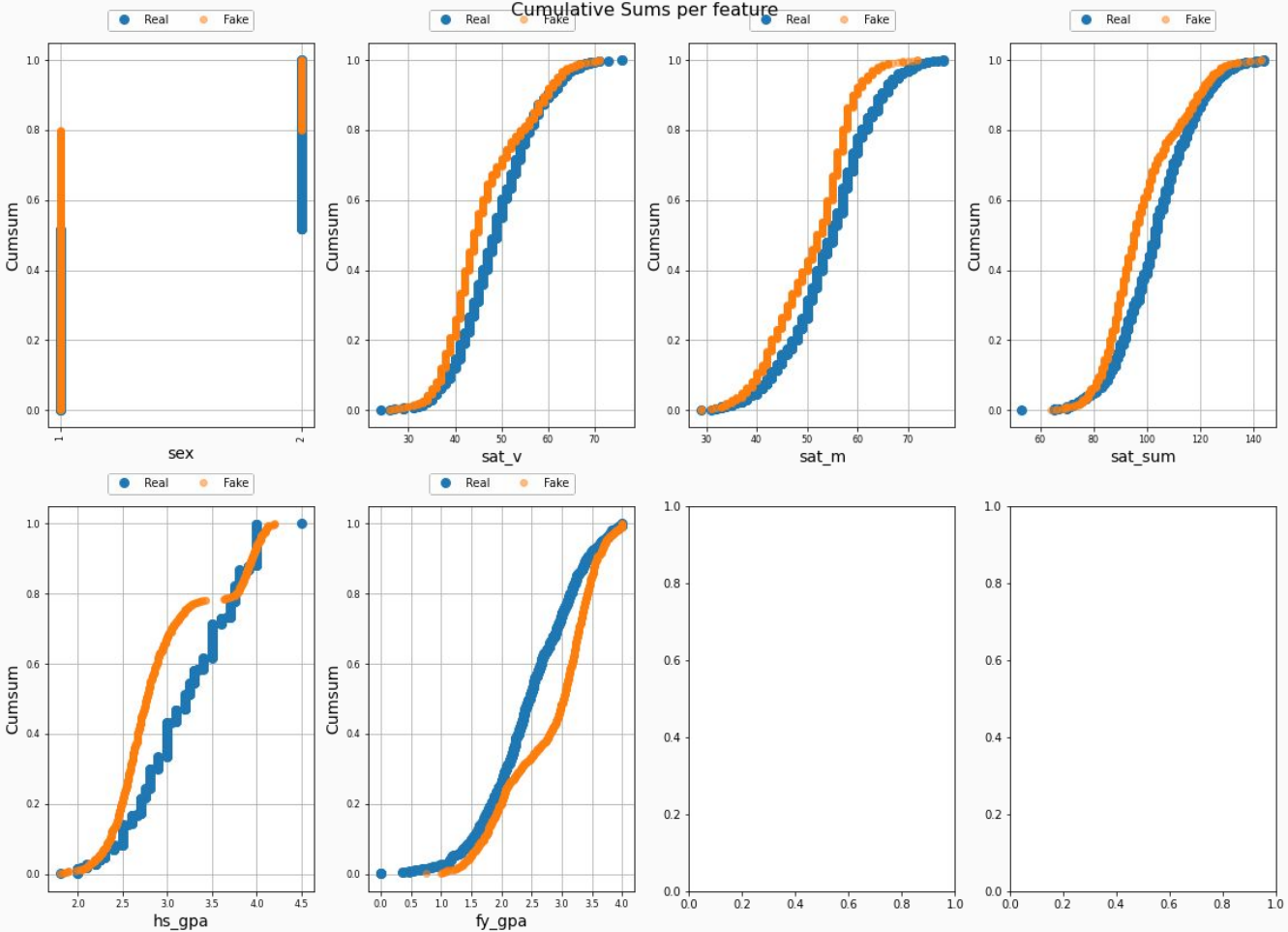
```
$no.uniques
[1] 1000

$no.replications
[1] 0

$per.replications
[1] 0
```

Apparent Match Distribution

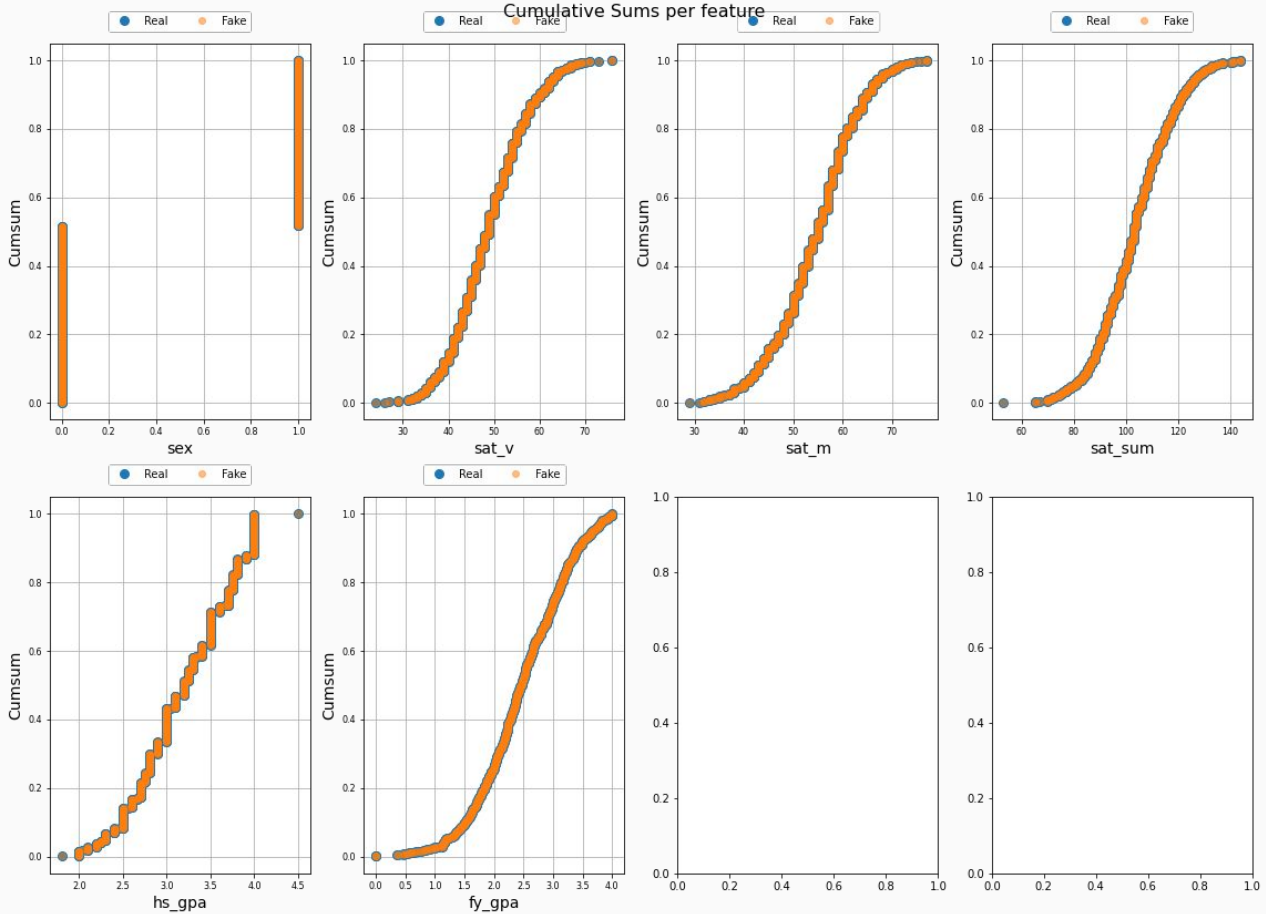
Record similarity distribution between pairs of apparently matching records, using partially synthesis of satgpa data. There is no apparent unique matches between the real and synthetic data.




Cumulative Sums per feature
Real
Fake
Cumsum
sex
sat_v
sat_m
sat_sum
hs_gpa
fy_gpa



Real vs Synthetic


Cumulative Sums per feature
Real
Fake
Cumsum
sex
sat_v
sat_m
sat_sum
hs_gpa
fy_gpa

Real vs Real

EDA~ ACS data: Help!

- Large data takes time for synthesize
  - Maybe divide and conquer!
  - Clustering
- Attributes: DEPARTS and ARRIVES
  - ARRIVES > DEPARTS
  - Possibly we need to convert it to minutes!
- Impose constraints
- Quasi identifiable attributes
  - Maybe: sex, age, marst, race, hispan, educ, citizen
- Sensitive attributes:
  - GQ, HCOVANY, HCOVPRIV, HINSEMP, HINSCAID, HINSCARE, EMPSTAT, EMPSTATD, LABFORCE, WRKLSTWK, ABSENT, LOOKING, AVAILBLE, WRKRECAL, WORKEDYR

For the presentation:

- Which methods we used and why
- Comparison of results using different methods
  - Utility
  - Privacy
- Interpretation of the results:
  - When is the utility ‘ good enough’ for the different use cases
  - When is the privacy ‘ good enough’ for the different use cases
- What would we recommend to management
- Evaluation of the guidebook
  - What worked well
  - What is missing or needs improving