Notes on our SAT-GPA Synthetic Datasets

Smart Data Foundry[1]
(At time of submission, GOFCoE)

24–28 January 2022
Version 1.1

Contents

[1] https://smartdatafoundry.com

1 SAT-GPA dataset

This dataset contains SAT (United States Standardized university Admissions Test) and GPA (university Grade Point Average) data for 1000 students at an unnamed college. Educational Testing Service originally collected the data. It can be found here.

1.1 Metadata

Field	Description	Sensitive	Type	Range	To be synthetize	Confidential/Dependent
sex	Gender of the student	No	Categorical	1,2	Yes	No
sat_v	Verbal SAT percentile	Yes	Continuous	[0,100]	Yes	No
sat_m	Math SAT percentile	Yes	Continuous	[0,100]	Yes	No
sat_sum	Total of verbal and math SAT percentiles	Yes	Continuous	[0,200]	No	No
hs_gpa	High school grade point average	Yes	Continuous	[0,5]	Yes	No
fy_gpa	First year (college) grade point average	Yes	Continuous	[0,5]	Yes	Yes

1.2 Data preparation

The `sat_sum` column can be obtained adding up `sat_v` and `sat_m`, and the sum is consistent with the two components for every record in the real SAT data, so it was removed before applying the models.

The column was calculated after the rest synthetic data was produced.

1.3 Data exploration

As part of data exploration we used the Miró implementation of test-driven data analysis (TDDA) to generate constraints describing the input data. These constraints specify the names, types minimum and maximum values for each field, and can (though in this case didn't) also generate constraints on uniqueness of values in a field, whether missing values are allowed. For strings fields, it can also generate regular expression constraints or lists of allowed values, though again these were not relevant here.

We also looked generate a new field and generate constraints for that. (It should

```
sat_sum_less_v_less_m = sat_sum - sat_v - sat_m
```

always be zero.)

A white paper describing automatic constraint generation and verification of data using the constraints is available as Automatic Constraint Generation and Verification.

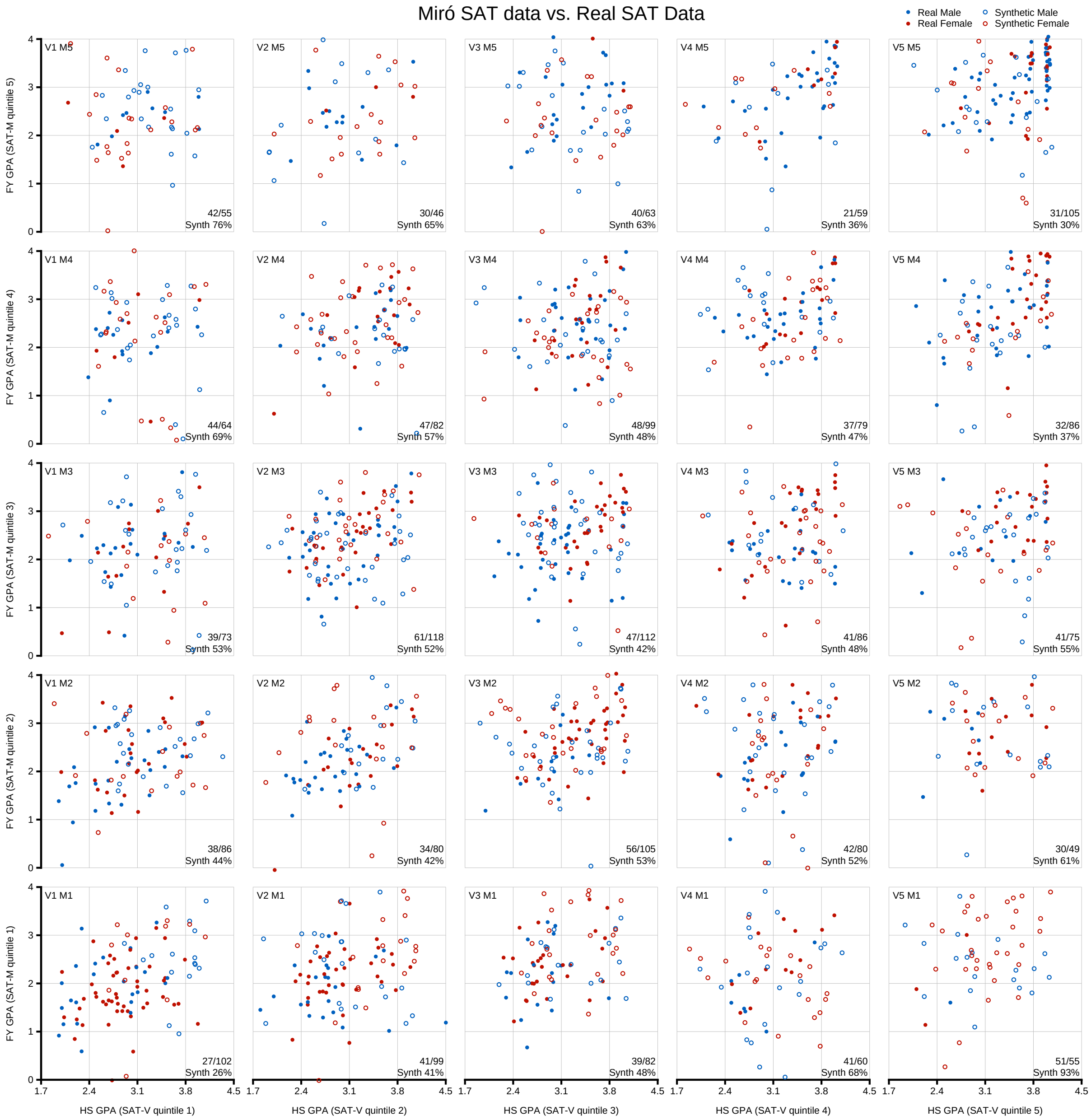
1.4 Explanation of Visualizations

The fact that the SAT data is so small (both in number of points, and dimensionality) makes is tempting to try to visualize *all* the data, in all the dimensions simultaneously. Indeed, we might also seek to compare real and synthetic data in a single visualization.

The SAT-GPA data has six dimensions (three SAT scores, two GPA scores and sex), but one of the SAT scores (`sat_sum`) is simply the sum of the others, so is linearly dependent and can be discarded. So we have five interesting dimensions in each dataset, of which four are numeric and one is two categories. If we compare a real to a synthetic dataset, we effectively have another binary/categorical variable.

We can obviously easily do scatter plots of any pair of the numeric variables. And we can certainly use colour to distinguish male and female. We could use two other colours and combine the categories e.g. (real male blue, real female red, synthetic male green, synthetic female orange), but it seems clearer to use different shapes, so we have chosen to use filled disks for real data and (empty) circles for synthetic data, using the same colours for real and synthetic.

We further realised, however, we could use small multiples to show discretized (binned) versions of two further numeric variables. This is illustrated below.



ZOOM FOR DETAIL

In this plot (which is a vector graph, so can be zoomed):

- each of the 25 scatter plots represents a combination of an approximate quintile for each of the two SAT Scores, with verbal SAT quintiles placed horizontally (major X) and math SAT quintiles place vertically (major Y). So
  - bottom left scatter plot is the lowest quintile for both `sat_v` and `sat_m`;
  - bottom right scatter plot is the *highest* quintile for `sat_v` and *lowest* quintile for `sat_m`;
  - top left is the *lowest* quintile for `sat_v` and *highest* quintile for `sat_m`;
  - top right is the highest quintile for both `sat_v` and `sat_m`.
- Red points are male (`sex=1`), blue points are female (`sex=2`).
- Solid points are real (and same on all plots), empty points are synthetic.
- Each subplot is scattergram of `FY_GPA` against `HS_GPA` against.

We have additionally annotated each plot with the number of synthetic points, N_S, and the total number of points, N, (synthetic + real) in the plot as N_S/N. Below that, we show that ratio as a percentage. So where the percentage is materially over 50%, there is an excess of synthetic data, and wher it is materially less, there is too little synthetic data.

The top-left annotation on each plot exicitly states the verbal and mathematical SAT quintiles (e.g. V1M5 for top left).

A very small amount of jittering has been applied to make points easier to distinguish where the data is dense.

2 Synthesis description

We have produced six different synthetic datasets from the original *SAT-GPA dataset*, using four different methods of synthesis and different levels of tuning:

- Miró (Stochastic Solutions)
- GEMINAI (Diveplane)
- Synthetic Data Vault using Copulas functions and Conditional Tabular GAN
- Synthpop with default parameters and the smooth setting.

2.1 Miró (Stochastic Solutions)

2.1.1 Discussion

Miró is a less well-known commercial analytics package produced by Stochastic Solutions Limited[2] and licensed by GOFCoE. It is written in Python and makes heavy use of `numpy`. When Miró generates constraints (in a `.tdda` file) using its version of TDDA, it can also write out distribution information (based on binning each field appropriately). Miró then has the ability to generate data that is consistent (or, in a few cases, largely consistent) with both the constraints and distributions saved in the TDDA file. In most cases, fields are treated entirely independently, i.e. by default Miró makes *no attempt* to generate data that preserves any relationships between fields. It is mostly used to generate high-quality test data that respects many of the constraints of the original data, rather than rich synthetic data that preserves correlations and other inter-field relationships.

- **Synthetic data:** `sat-synthetic-miro.csv`
- **Profile:** Synthetic Miró SAT Data

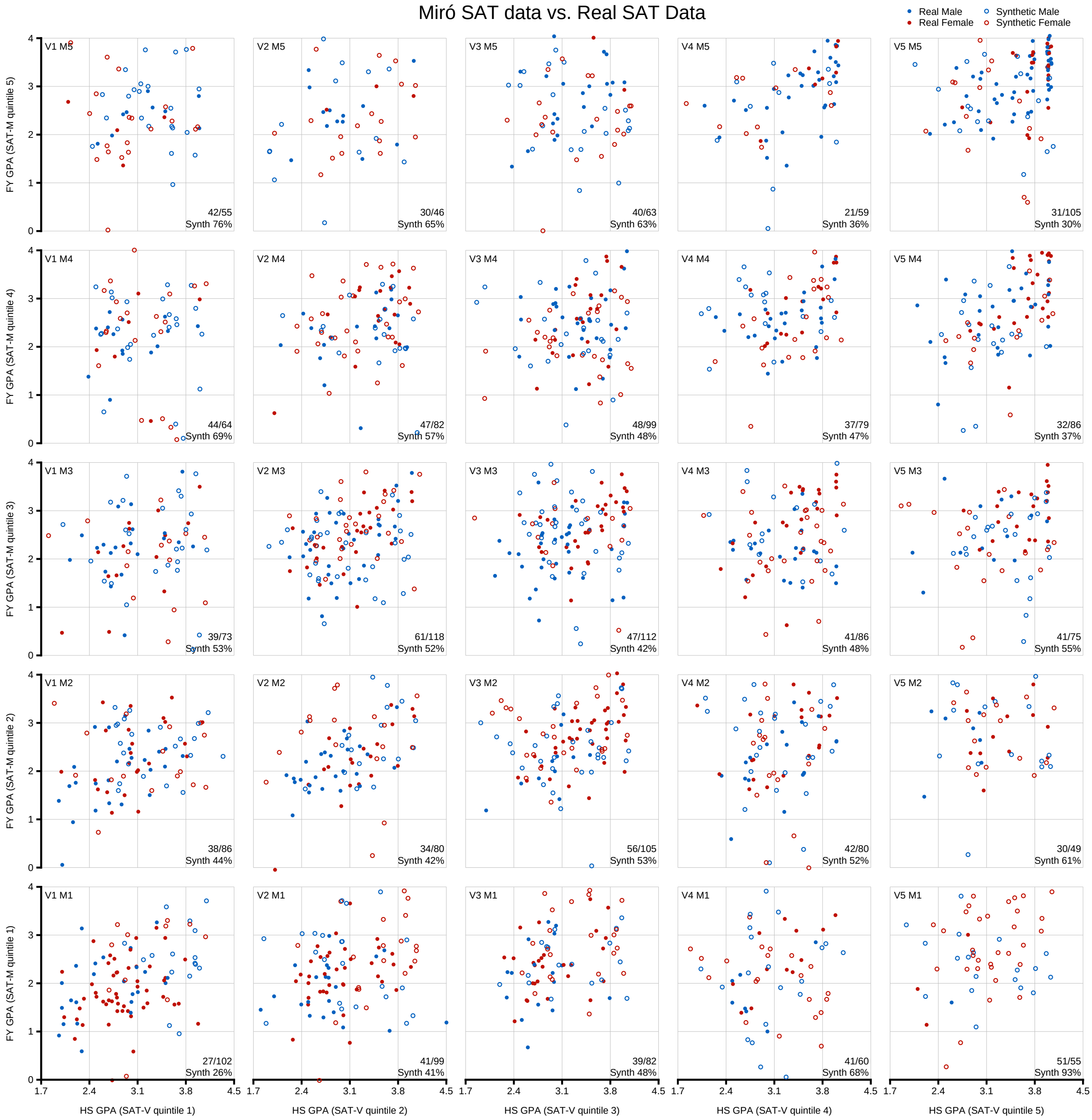
2.1.2 Visualization and Discussion (Miró Synthesis)

Miró synthesizes each dimension independently using only information about the distribution of that field. So we should expect very little relationship between the real and synthetic data when looking in all five dimensions as here.

If you zoom into the plot you see this. Note particularly:

- Because the real data has strong positive correlation between the two SAT scores, Miró has far more points in the off-diagonal plots than does the real data. Accordingly, Miró's synthetic data accounts for 76% of the data in the top-left scatter plot and 93% of the data in the bottom-right scatter plot.
- On the leading diagonal, the distributions within the plots are very different for the real and synthetic data. For example, the bottom left plot (low SAT scores) shows most of the real data in the bottom left of the sacttergram, whereas most of the synthetic data is in the upper right. Conversely, in the upper right scattergram, most of the real data is concentrated at a High-School GPA very close to 4.0, and a math FY GPA between 3.0 and 4.0, whereas the synthetic data, while also mostly in the top right, shows no correlation between the two GPA scores.
- From a disclosure risk perspective, there is no real tendency for the synthetic points to "shadow" real points: there occasional close real and synthetic points of the same colour (e.g. V3M5 at about (2.5, 3.3), but that plausibly (and actually) arises largely by chance.

[2] One of the team members, Nick Radcliffe, is a principal author of Miró, and founder director of Stochastic Solutions Limited



ZOOM FOR DETAIL

2.2 GEMINAI (Diveplane)

2.2.1 Discussion

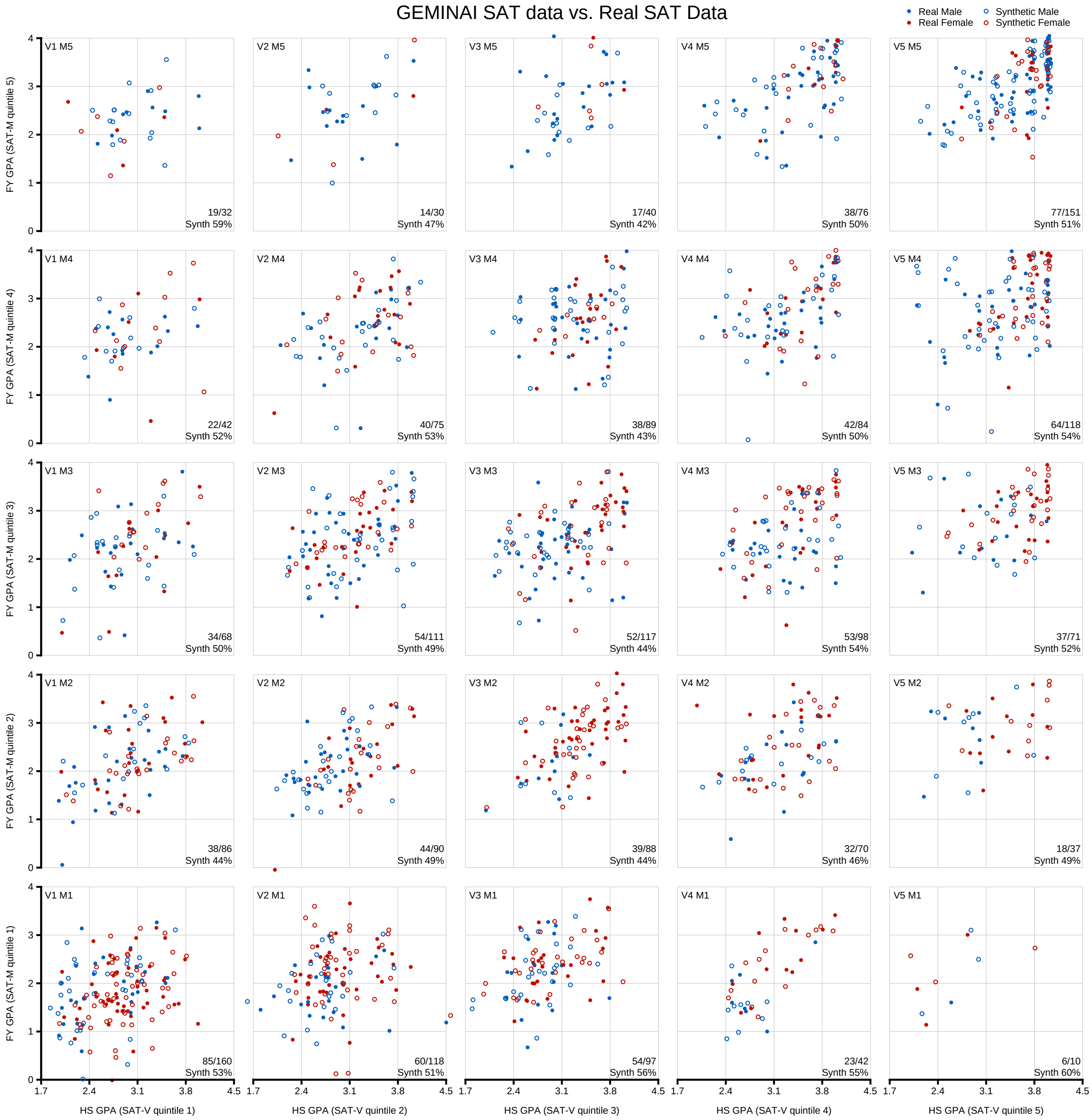
GEMINAI [1] uses a proprietary synthesis method based on k-Nearest Neighbours and Differential Privacy, and provides a broad suite of assessment metrics covering both utility and disclosure risk, which we have used across the various datasets we have synthesized using all methods.

- **Synthetic data:** `sat-synthetic-diveplane.csv`
- **Profile:** Synthetic GEMINAI SAT Data

2.2.2 Visualization and Discussion (GEMINAI)

This visualisation shows that in general the GEMINAI synthesizer produced convincingly similar distributions to the real data across all dimensions. However, there are synthetic and real points in the same class (Male or Female) that do look too close to each other for comfort, which might create a disclosure risk problem.

Notice also that in every scatterplot the proportion of the data that is synthetic is reasonably close to 50%, with the highest being 60% (bottom right) and lowest being 42% (middle top).



ZOOM FOR DETAIL

2.3 Synthetic Data Vault Gaussuan Copula (SDV-CG)

Gaussian Copula: As described in their documentation SDV Gaussian Copulas

A copula is a distribution over the unit cube $[0, 1]^d$ *which is constructed from a multivariate normal distribution over* $\mathbb{R}^d$ *by using the probability integral transform. Intuitively, a copula is a mathematical function that allows us to describe the joint distribution of multiple random variables by analysing the dependencies between their marginal distributions.*

```
import pandas as pd
from sdv.tabular import GaussianCopula

def main():
    data = pd.read_csv("satgpa.csv")
    model = GaussianCopula()
    model.fit(data)
    model.save(satgpa_gcopula.pkl)

if __name__ == "__main__":
    main()
```

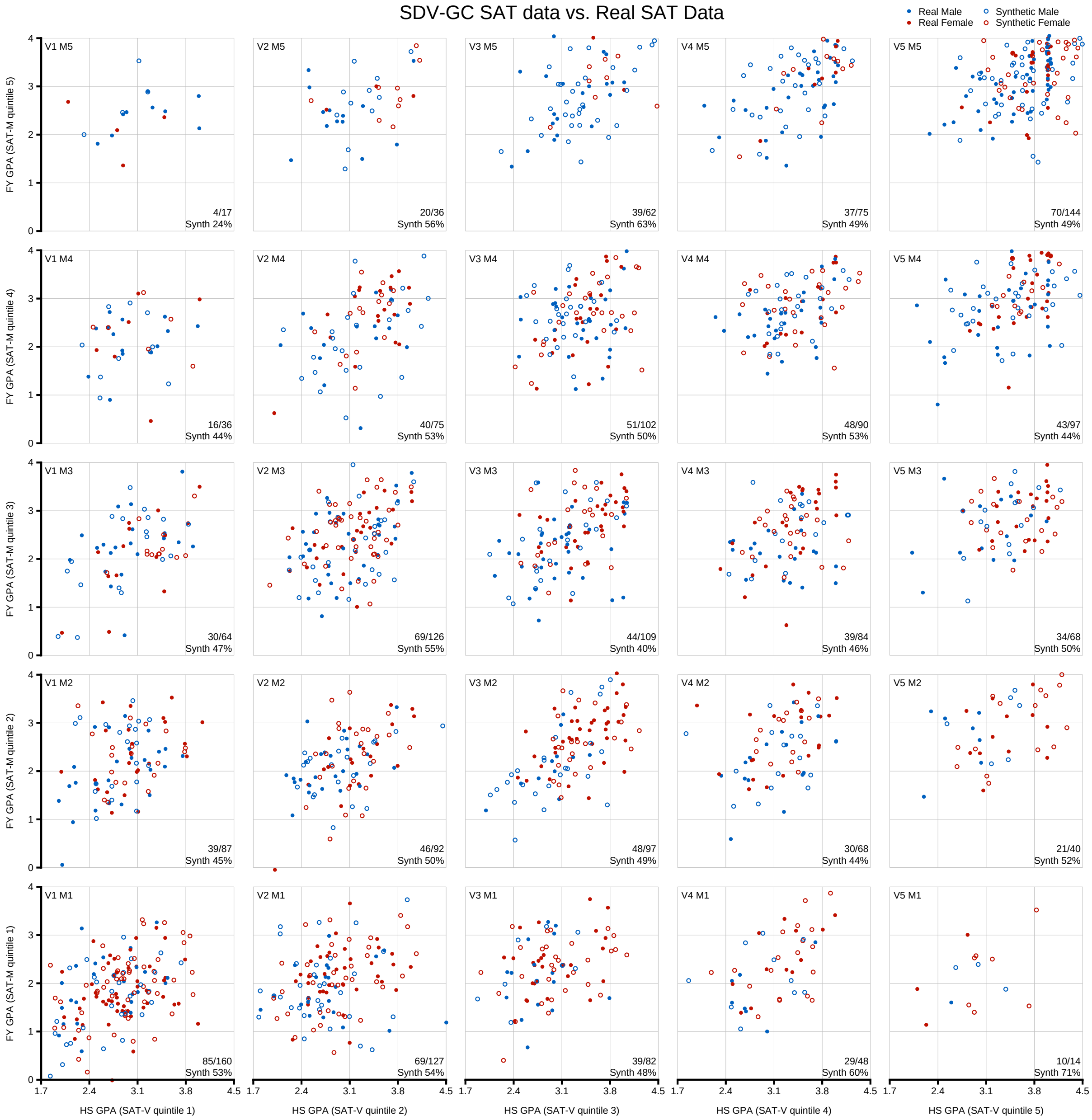
- **Synthetic data:** `sat-synthetic-sdv-gcopula.csv`
- **Profile:** Synthetic SDV-GC SAT Data

2.3.1 Visualization and Discussion (SDV-GC Synthesis)

Overall, the SDV Gaussian Copula synthesizer did a very convincing synthesis. However, it is clear from the graphs that in the top and bottom-most quintiles for each of the SAT scores, (the corner scatterplots) it had more difficulty re-creating the same full structure. A clear example is the upper right scatter plot where the line of real points was not re-created for the synthetic data, which was a bit more scattered. Also, we can single out data points that might have a higher privacy risk.

The proportions of real and synthetic data do vary significant across the scattergrams, with on 24% of the points in the upper left plot being synthetic, and 71% of the points being synthetic in the bottom right (though that only amounts to 10/14 points).

Notice also that SDV-GC has failed to produce any females in the top left scattergram, i.e. with the highest SAT scores in mathematics but the lowest for verbal skills.



ZOOM FOR DETAIL

2.4 Synthetic Data Vault Conditional Tabular GAN (SDV-CTGAN)

2.4.1 Discussion

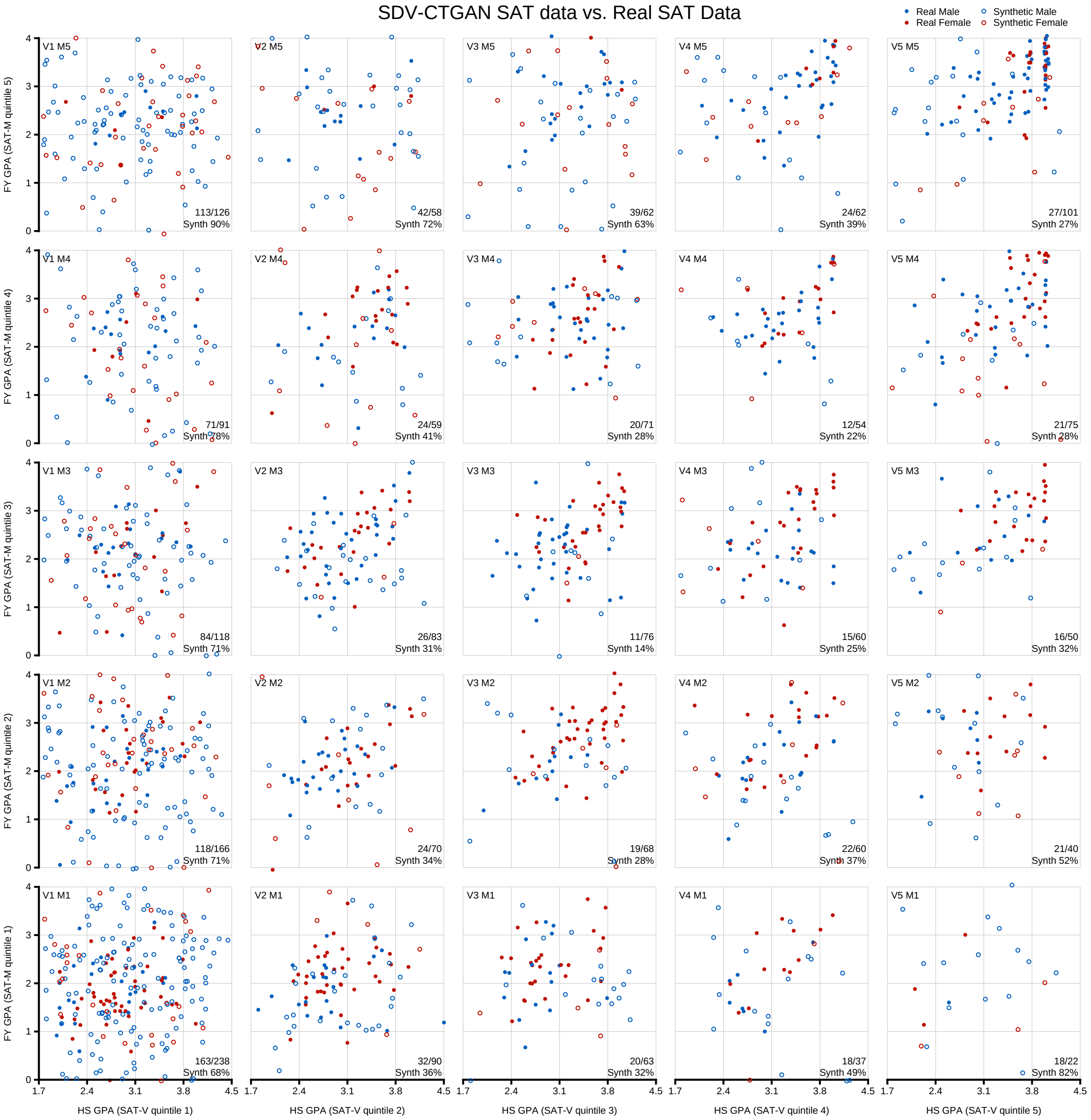
CTGAN: Short for *Conditional Tabular GAN* is a GAN-based deep learning model which was presented at the NeurIPS 2020 conference [2]. We didn't set any primary id key for this dataset because there is no unique id of interest.

```
import pandas as pd
from sdv.tabular import CTGAN

def main():
    data = pd.read_csv("satgpa.csv")
    model = CTGAN()
    model.fit(data)
    model.save(satgpa_ctgan.pkl)

if __name__ == "__main__":
    main()
```

- **Synthetic data:** `sat-synthetic-sdv-ctgan.csv`
- **Profile:** Synthetic SDV-CTGAN Data

2.4.2 Visualization and Discussion (SDV-CTGAN Synthesis)

In general, the SDV-CTGAN dataset was the one with the lowest quality in terms of utility. We can see that the proportions of synthetic data points in the first `sat_v` quantile it did over-shoot the amount of real data, whereas the higher `sat_v` scores it had too little data by 10-30 percentage points. This method, being a deep-learning based approach, might require more data to work well.



ZOOM FOR DETAIL

2.5 Synthpop-Default (University of Edinburgh)

2.5.1 Description

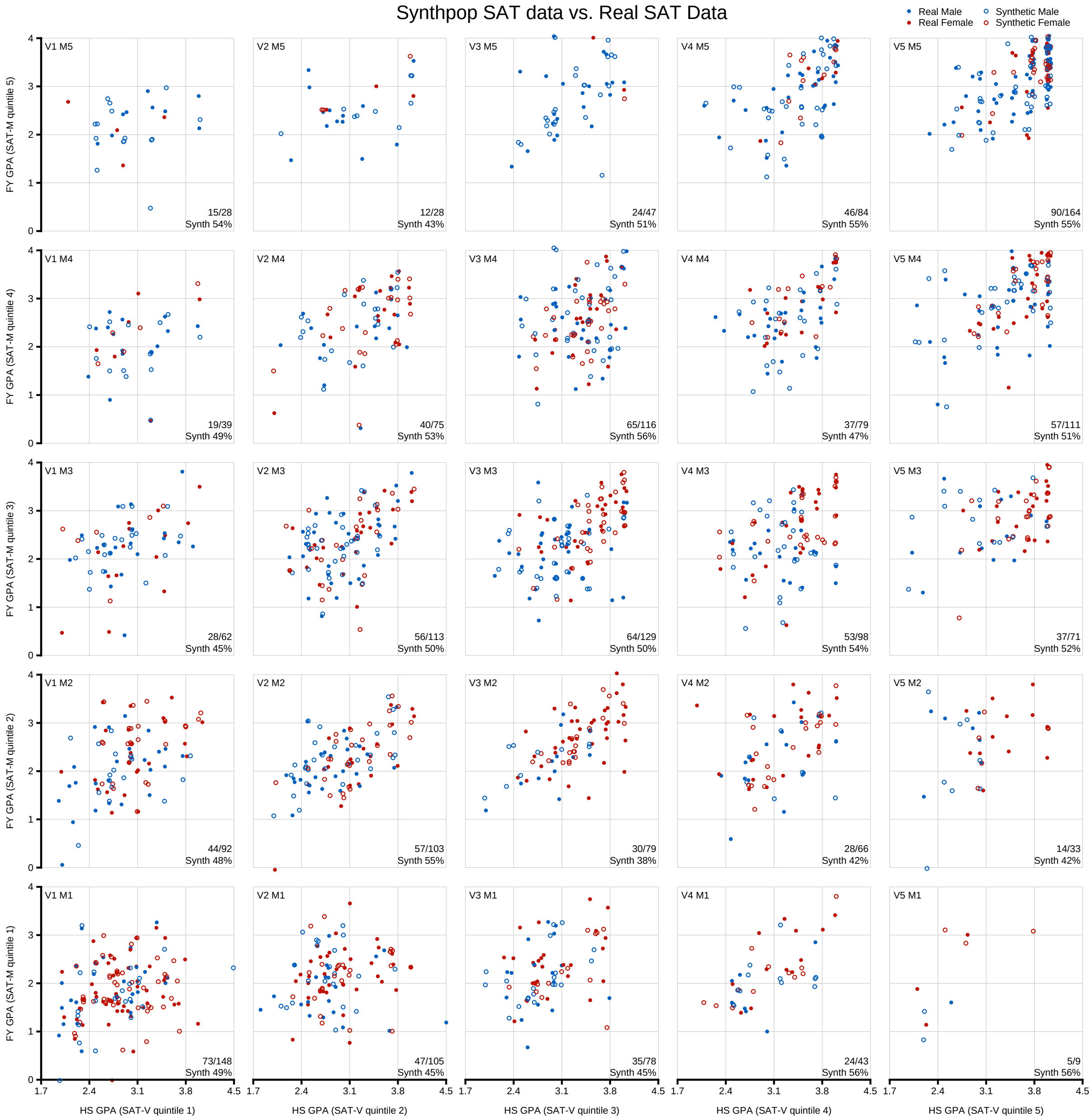
Synthpop provides default parameters to synthetise the data:

- CART method used (except for the first variable)
- Synthesis incremental in the variables order presented in the dataset
- All previously synthesised variables used as predictors
- The `sex` variable was converted to factors
- **Synthetic data:** `sat-synthetic-synthpop.csv`
- **Profile:** Synthetic Synthpop-Default SAT Data

2.5.2 Visualization and Discussion (Synthpop-Default Synthesis)

Synthpop using the default parameters did one of the best syntheses. As we see on the upper right quantile, it convincingly mimicked the point distribution. However, one of the bigger concern of synthpop is the disclosure risk. Many of the synthetic points were too close to the real points for comfort, highlighting the extra care that we need too have examining the Utility-Privacy trade-off.

As with SDV-GC, Synthpop here failed to produce any females in the top left scattergram, i.e. with the highest SAT scores in mathematics but the lowest for verbal skills.



ZOOM FOR DETAIL

2.6 Synthpop-Smooth (University of Edinburgh)

2.6.1 Description

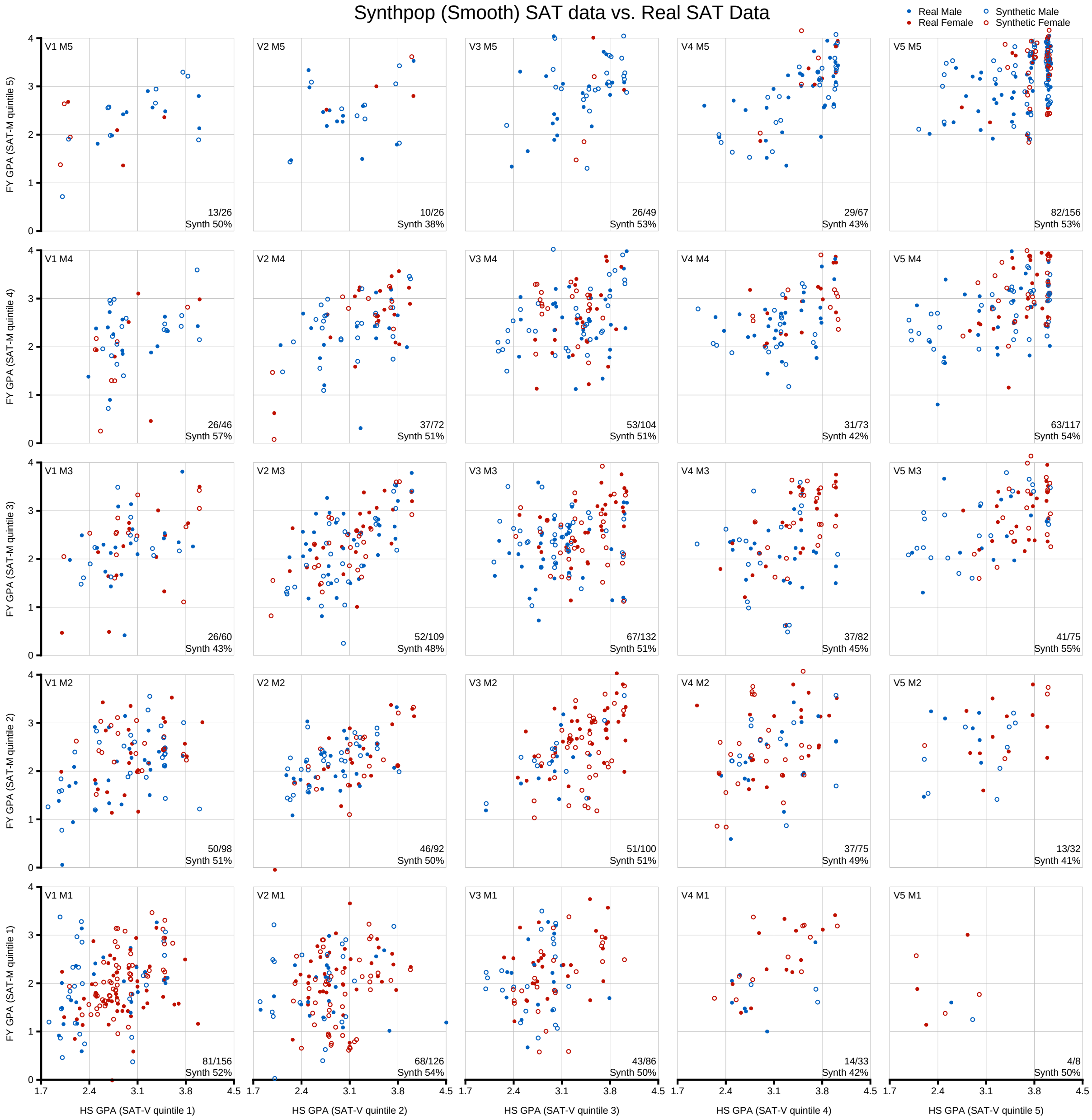
Additionally, an alternative version was produced called **"Synthpop Smooth"**. The main changes introduced were:

- hs_gpa column moved to be second in the order.
- Smoothing option was activated for the variables: sat_v, sat_m and fy_gpa.
- Option proper was set to TRUE. More info here.
- **Synthetic data:** sat-synthetic-synthpop-smooth.csv
- **Profile:** Synthetic Synthpop-Smooth SAT Data

2.6.2 Visualization and Discussion (Synthpop-Smooth Synthesis)

Synthpop-smooth did even better in terms of producing a very similar multidimensional synthetic double of the real SAT-GPA data. Notice that in this case it did, for example, produce females in the top left plot, and again matched the dense stripe in the uppoer right well.

Perhaps even more than SDV with default parameters, there is some concern that some synthetic points were very close to real points. ¡



ZOOM FOR DETAIL

References

[1] Christopher J. Hazard, Christopher Fusting, Michael Resnick, Michael Auerbach, Michael Meehan, & Valeri Korobov. “Natively Interpretable Machine Learning and Artificial Intelligence: Preliminary Results and Future Directions,” https://arxiv.org/abs/1901.00246.

[2] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante and Kalyan Veeramachaneni “Modeling Tabular data using Conditional GAN,” https://arxiv.org/abs/1907.00503.

[3] Nowok, Beata and Raab, Gillian M. and Dibben Chris “synthpop: Bespoke Creation of Synthetic Data in R,” https://www.jstatsoft.org/index.php/jss/article/view/v074i11.