Higher Education Synthesized

Extending Higher Education Data with Synthetic Data at the Onset of the Demographic Cliff

Problem Statement — Introduction/Motivation

Data Mining + Diminishing Data = No Good

- After decades of enrollment increases, higher education is now facing decreasing enrollments, either gradually over time or with a sharp drop off a demographic cliff, depending on who you ask

- One corollary to declining enrollments is declining amounts of data

- This project will explore synthesizing institution-level Integrated Postsecondary Education Data System (IPEDS) data from the National Center for Education Statistics (NCES), a well-known and reliable set of higher education data overseen by the federal Department of Education

- This iteration of the project will use generative models in the Synthetic Data Vault Project (https://github.com/sdv-dev) to create tabular synthetic IPEDS data and evaluate the quality of that data

- Source code at https://github.com/jhleakakos/msds-data-mining-project

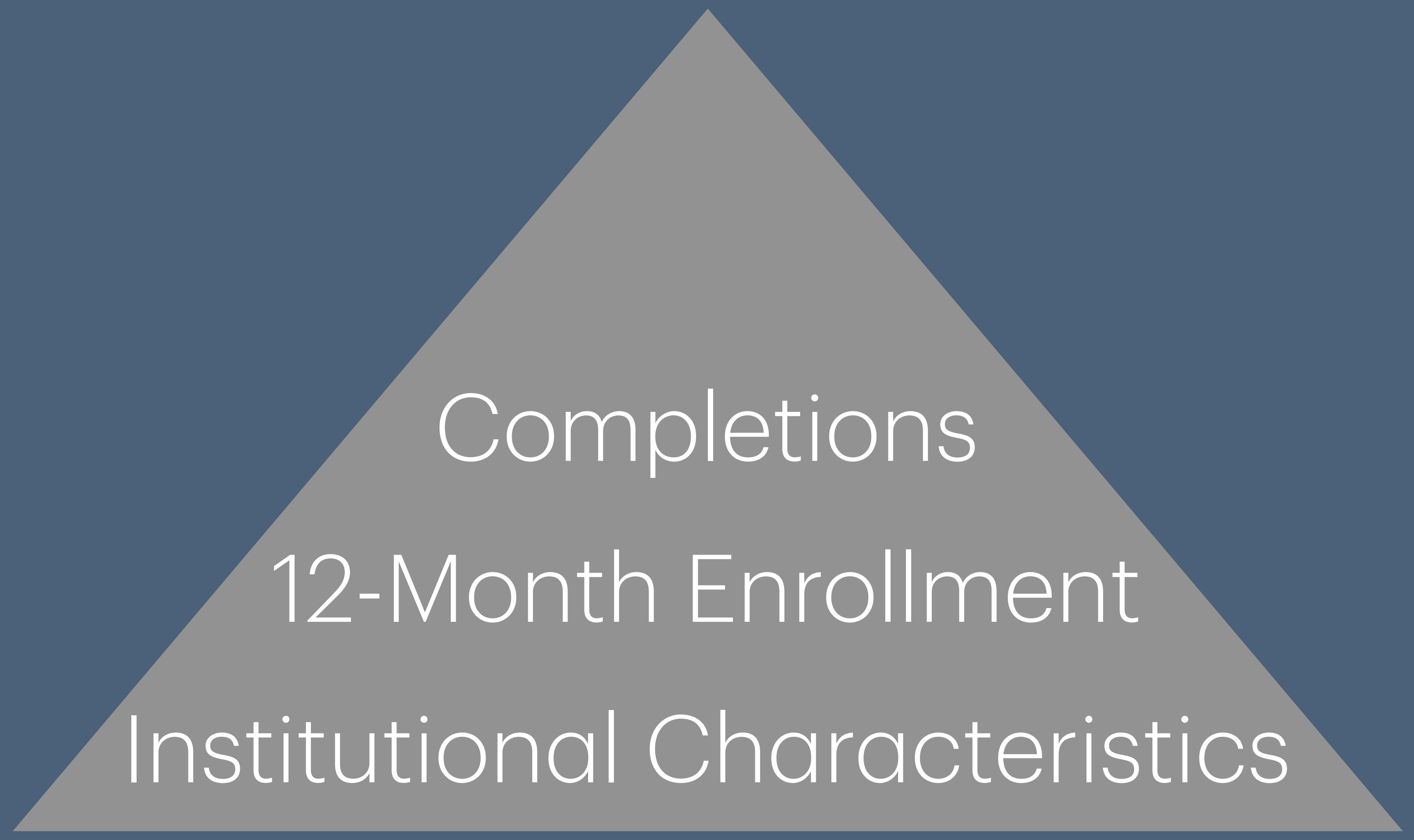
IPEDS Data

Domains of Data

Institutional Characteristics: information related to higher education institutions

12-Month Enrollment: unduplicated counts of students who enroll anytime during a 12-month period

Completions: counts of students who complete degree or non-degree credentials



Completions

12-Month Enrollment

Institutional Characteristics

Difficulties

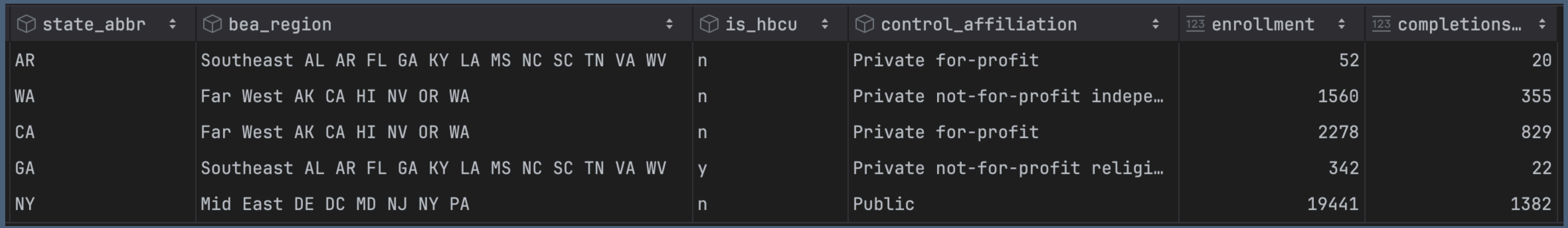
- IPEDS data dictionaries are not clear about mapping categorical feature numeric values to definitions
- IPEDS data dictionaries are not clear about how to group across categorical features to get the correct summary counts
- Number of features per data file — some files have hundreds of features
- Inconsistencies between encodings (binary fields represented as different pairs of numbers), even with the same files

Synthetic Data

- Synthetic data is “artificially generated [data] that resemble[s] the actual data — more precisely, having similar statistical properties” (*)

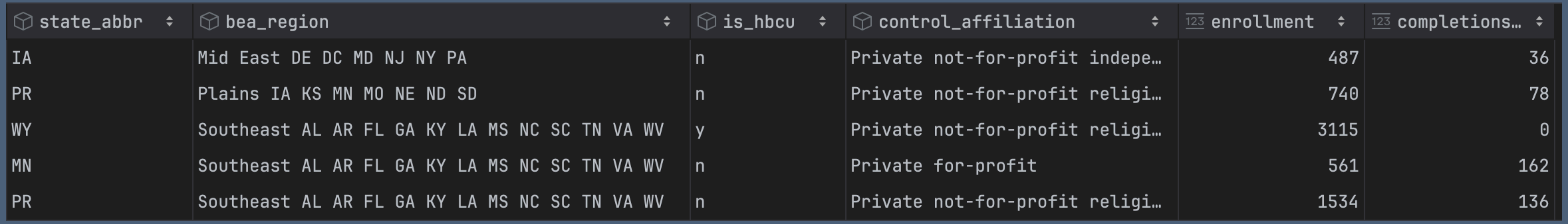
- Synthetic data allows us to create more IPEDS data to counter the reduction in real higher education data, either to replace or supplement that real data

Real Data



state_abbr	bea_region	is_hbcu	control_affiliation	enrollment	completions…
AR	Southeast AL AR FL GA KY LA MS NC SC TN VA WV	n	Private for-profit	52	20
WA	Far West AK CA HI NV OR WA	n	Private not-for-profit indepe…	1560	355
CA	Far West AK CA HI NV OR WA	n	Private for-profit	2278	829
GA	Southeast AL AR FL GA KY LA MS NC SC TN VA WV	y	Private not-for-profit religi…	342	22
NY	Mid East DE DC MD NJ NY PA	n	Public	19441	1382

Synthetic Data



state_abbr	bea_region	is_hbcu	control_affiliation	enrollment	completions…
IA	Mid East DE DC MD NJ NY PA	n	Private not-for-profit indepe…	487	36
PR	Plains IA KS MN MO NE ND SD	n	Private not-for-profit religi…	740	78
WY	Southeast AL AR FL GA KY LA MS NC SC TN VA WV	y	Private not-for-profit religi…	3115	0
MN	Southeast AL AR FL GA KY LA MS NC SC TN VA WV	n	Private for-profit	561	162
PR	Southeast AL AR FL GA KY LA MS NC SC TN VA WV	n	Private not-for-profit religi…	1534	136

* Emiliano De Cristofaro. 2024. Synthetic Data: Methods, Use Cases, and Risks. arXiv:2303.01230 [cs.CR] https://arxiv.org/abs/2303.01230

Related Work

We're On Our Own

- Did not find much research directly related to data mining or generative modeling for higher education or IPEDS data
- Found some research about data mining on education data generally, particularly focusing on different predictive approaches
- Found research speaking to the importance of bringing more advanced data methods to higher education data
- Found research highlighting the lack of big data practices in higher education
- Found a subsection of research focused on general use of synthetic data
- In summary, no research that covers each of the aspects of this project

Proposed Work

1. Scrape IPEDS data from the IPEDS website using Python scripts, storing flat files on disk

2. Leverage IPEDS data dictionaries to select fields that we want to explore for possible modeling

3. Run a local PostgreSQL database in Docker, using schemas to separate out data lake and data warehouse

4. Design, create, and load data lake with SQL scripts based on findings in data dictionaries

5. Pull data out of data lake and into Jupyter Notebook for an initial pass of EDA

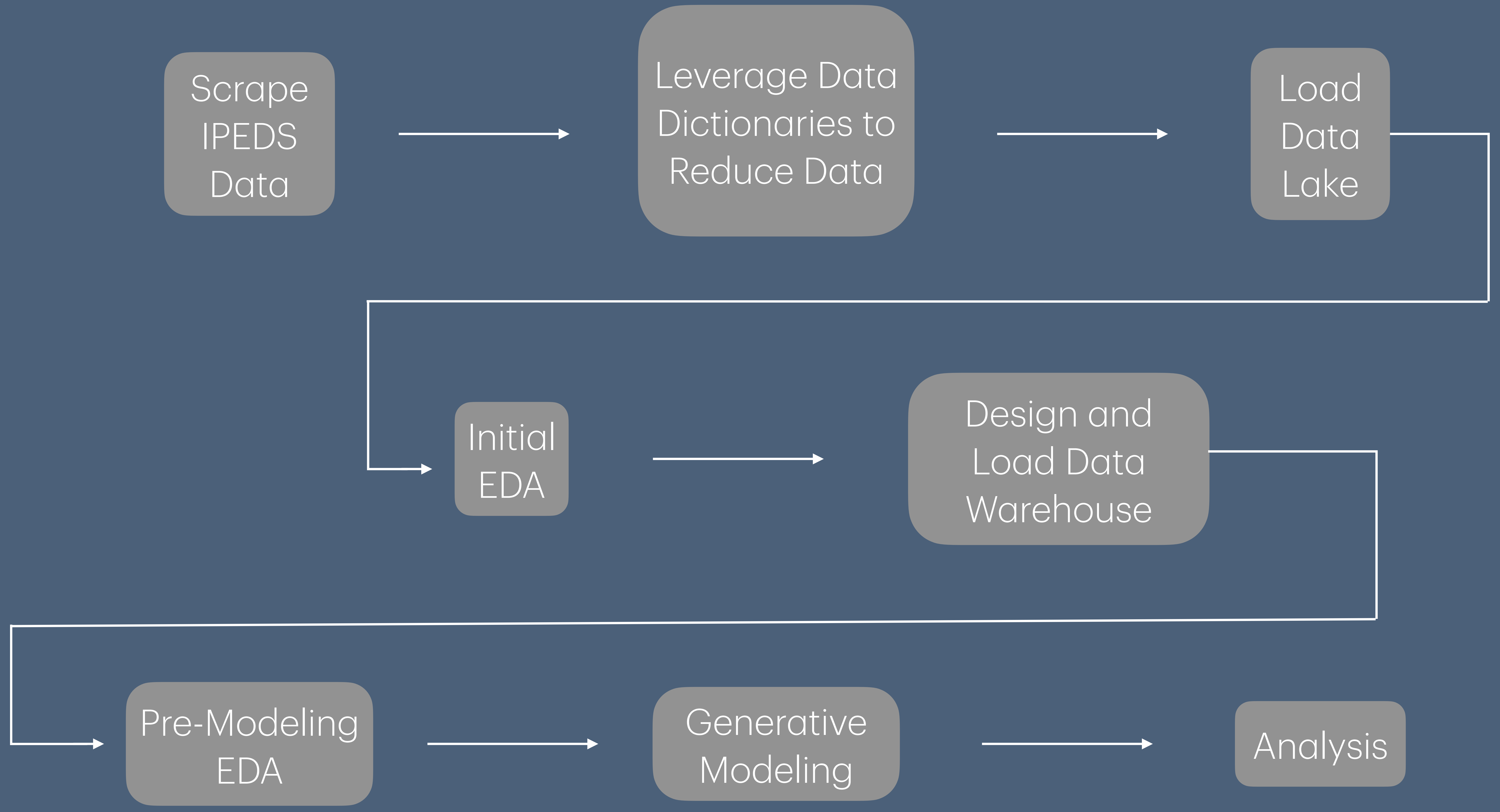
6. Use EDA to inform design of data warehouse, using a very small-scale version of a dimensional model to mimic a larger warehouse concept

7. Pull data out of data warehouse and into Jupyter notebook for generative modeling, using Synthetic Data Vault Project libraries to generate synthetic IPEDS data

8. Evaluate synthetic data quality

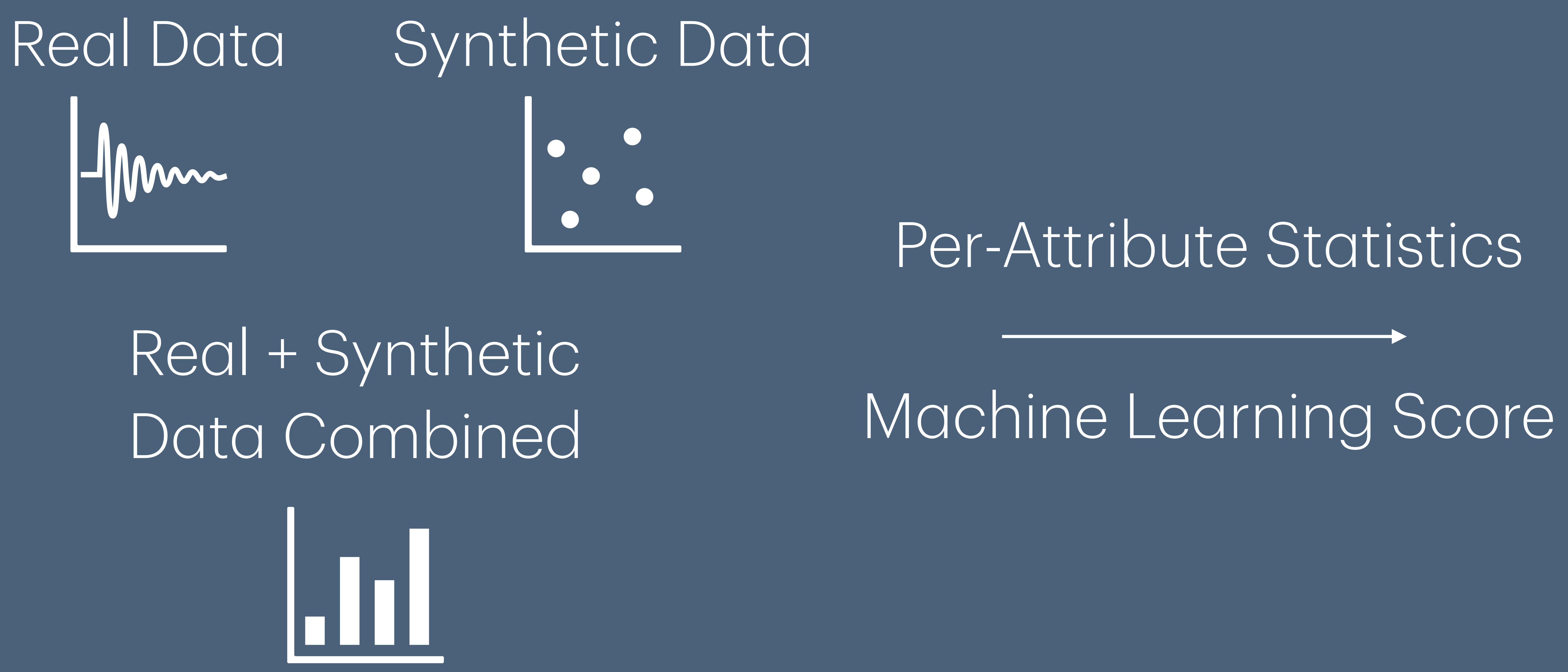
9. Load real and synthetic data into a separate Jupyter Notebook in order to perform deeper modeling and run modeling on real data, synthetic data, and a combination of both

Data Pipeline



Evaluation Methodology

Is Our Fake Data Any Good?



Real Data

Synthetic Data

Real + Synthetic Data Combined

Per-Attribute Statistics

Machine Learning Score

Data	Metric	Value
Real	Median	x.xx
Synthetic	Median	x.xx

Data	MSE	R^2
Real	x.xx	x.xx
Synthetic	x.xx	x.xx
Combined	x.xx	x.xx

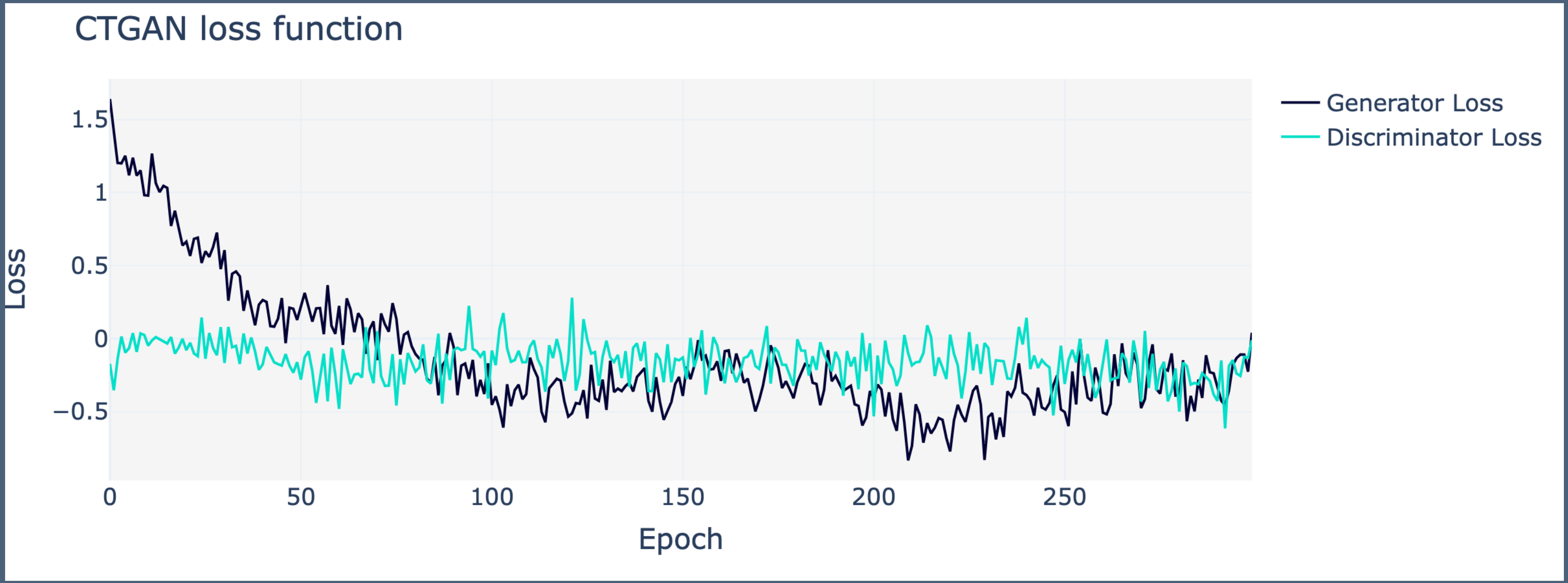
- **Per-Attribute Statistics**: comparing mean, median, standard deviation, and other distributional metrics between features in the real and synthetic data

- **Machine Learning Score**: training regressors on each of the three input training sets and compare performance

Evaluation Metrics

- Does the GAN training look reasonable in terms of generator and discriminator losses?
- Are measures like the mean and standard deviation for a given feature similar across the real and synthetic data?
- Do synthetic columns have the same distribution shapes as their associated real columns?
- Do trends between columns in the real data show the same trends in the synthetic data?
- Is the synthetic data valid — unique primary keys, values that fall within the minimum and maximum range for numeric columns, and categories that exist in the real data?
- How similar is the performance of regressors when trained on the real, synthetic, and combined data and tested on a held-out test set from the real data?

Evaluation Output



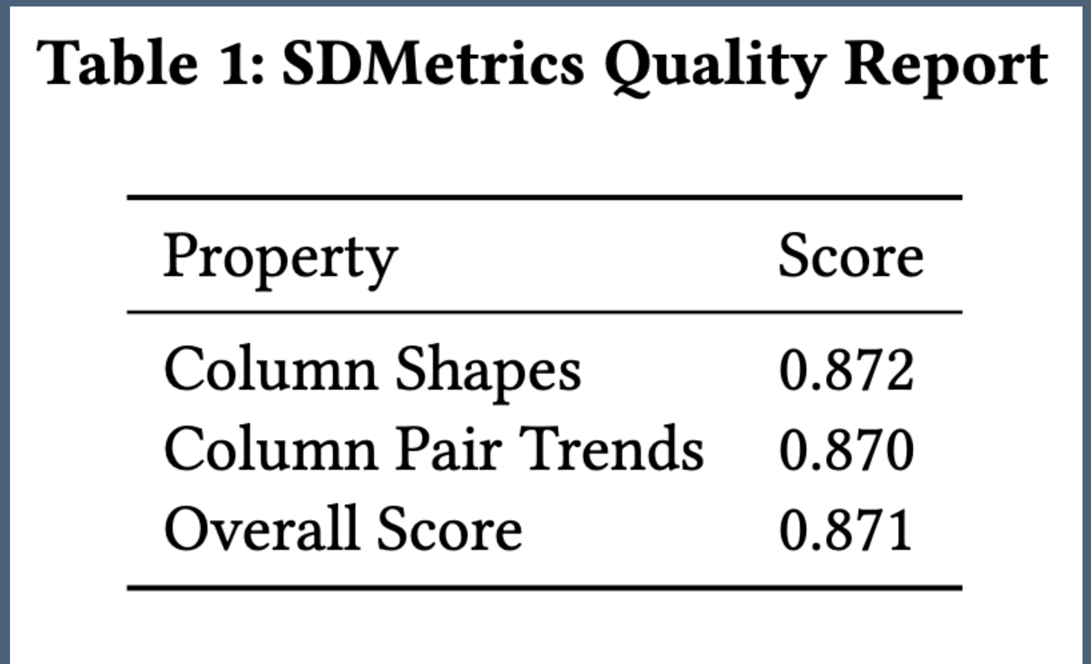


Table 1: SDMetrics Quality Report

Property	Score
Column Shapes	0.872
Column Pair Trends	0.870
Overall Score	0.871

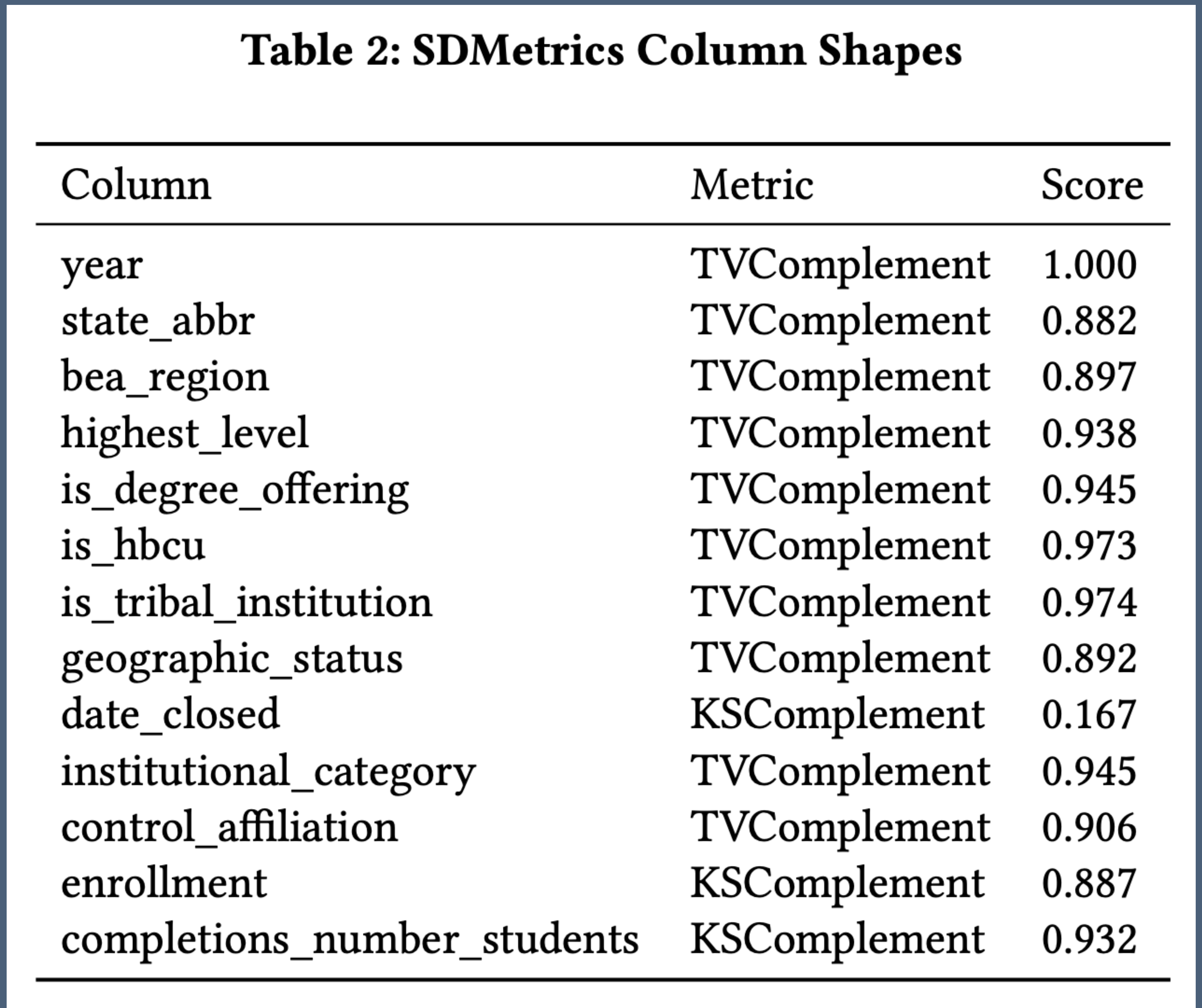


Table 2: SDMetrics Column Shapes

Column	Metric	Score
year	TVComplement	1.000
state_abbr	TVComplement	0.882
bea_region	TVComplement	0.897
highest_level	TVComplement	0.938
is_degree_offering	TVComplement	0.945
is_hbcu	TVComplement	0.973
is_tribal_institution	TVComplement	0.974
geographic_status	TVComplement	0.892
date_closed	KSComplement	0.167
institutional_category	TVComplement	0.945
control_affiliation	TVComplement	0.906
enrollment	KSComplement	0.887
completions_number_students	KSComplement	0.932

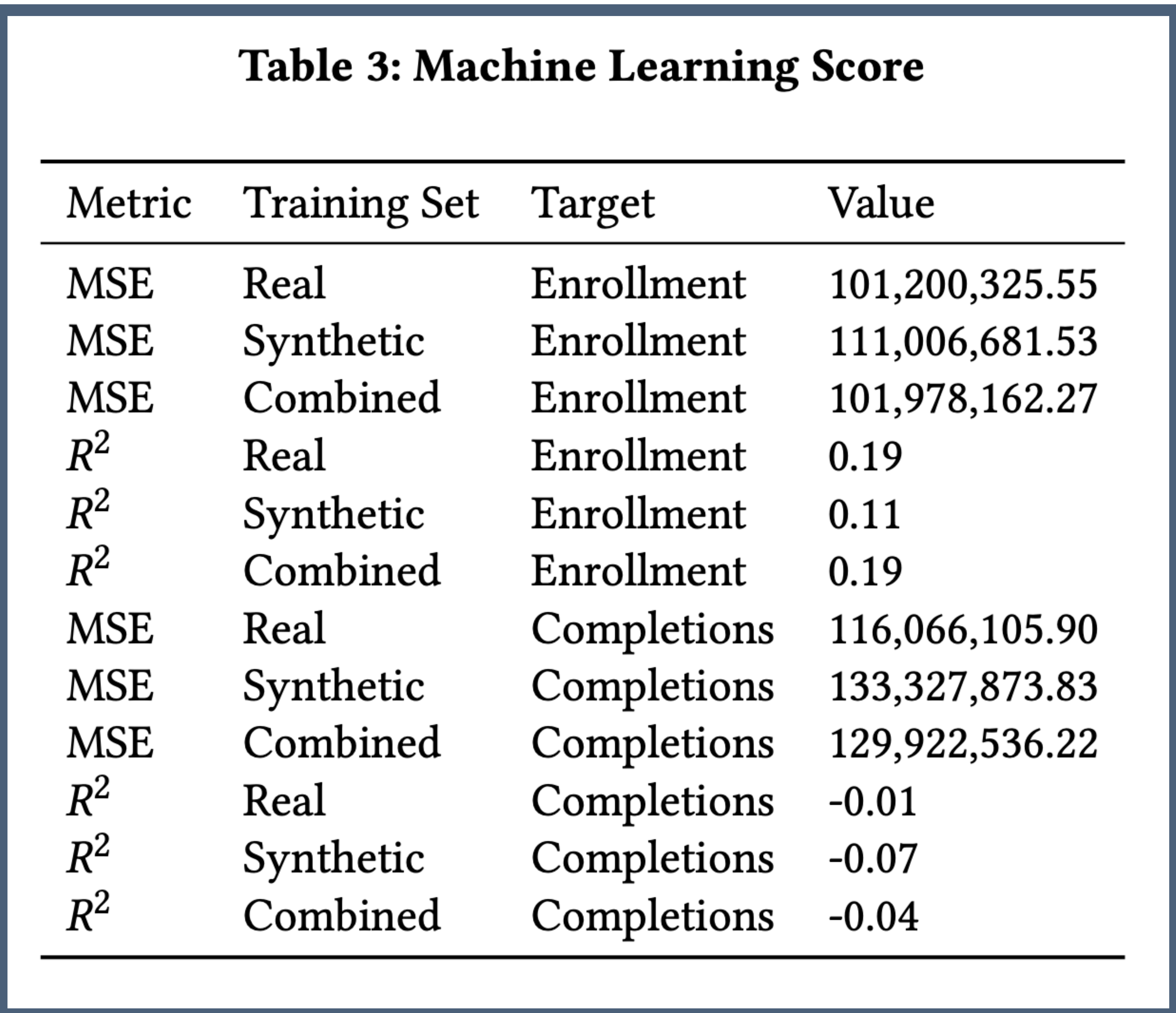
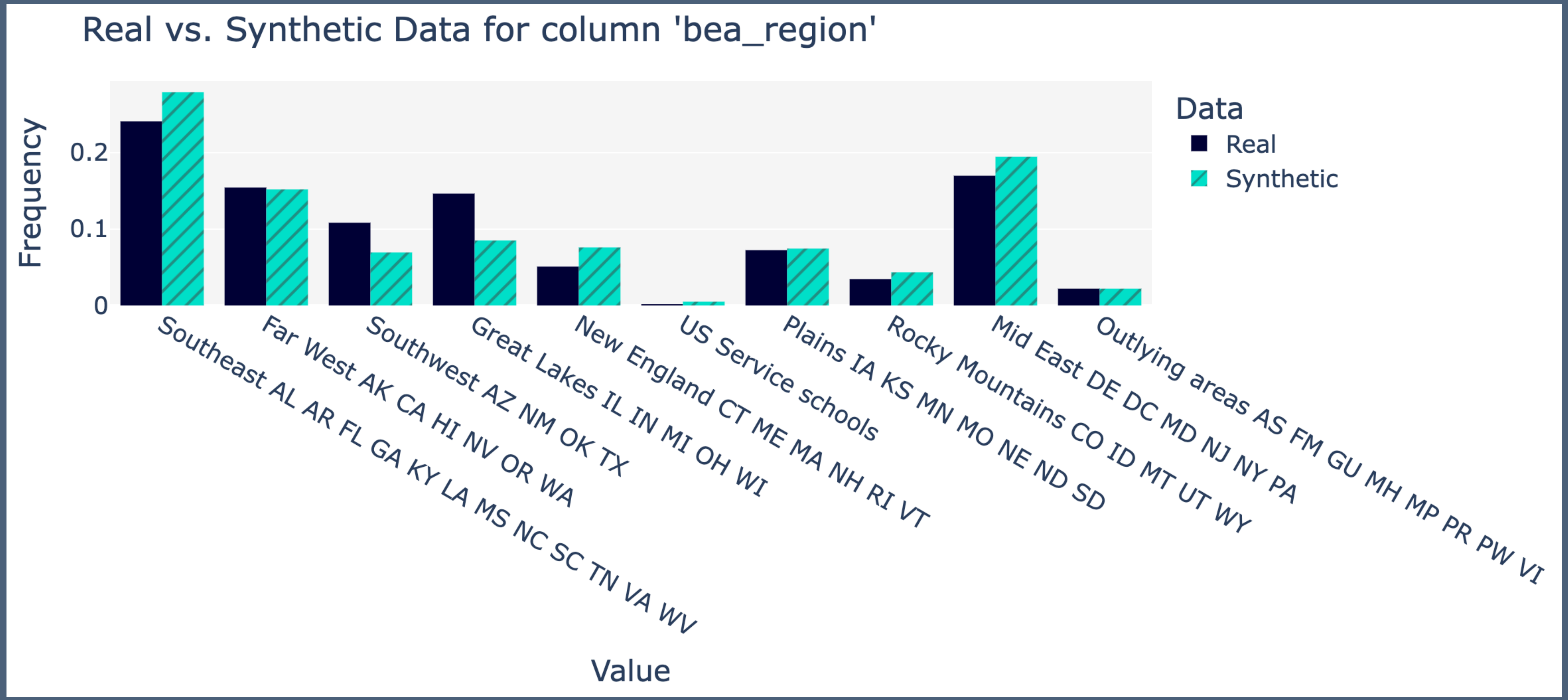
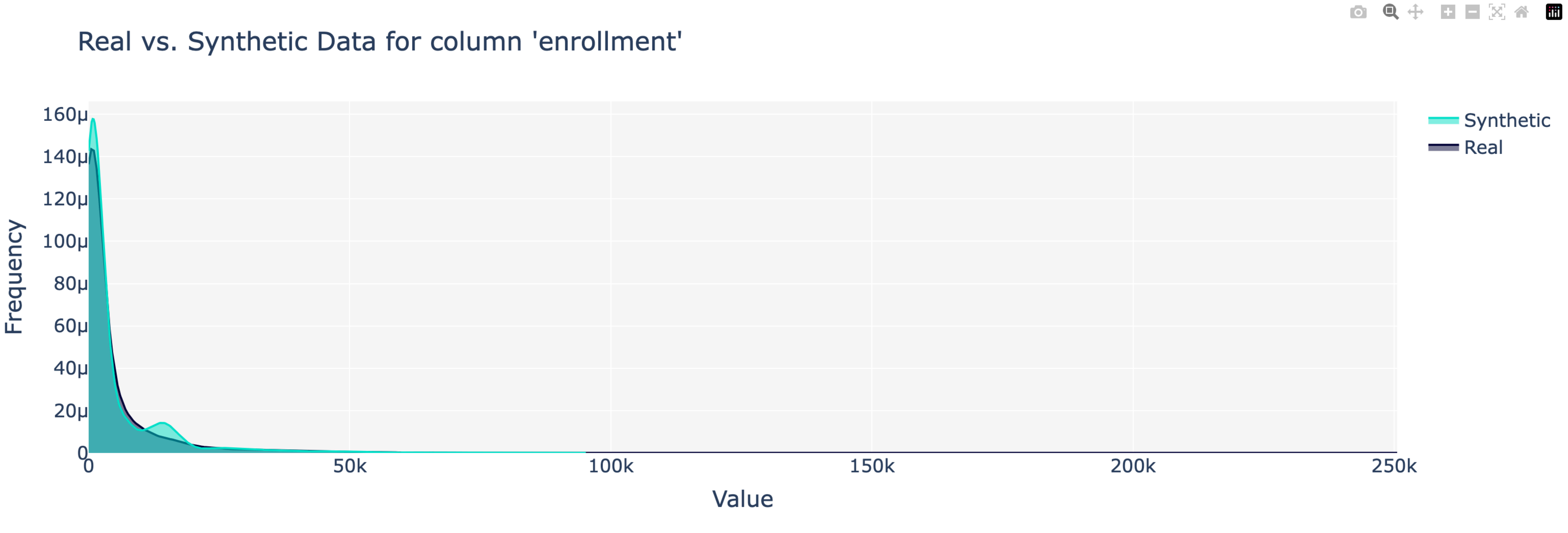


Table 3: Machine Learning Score

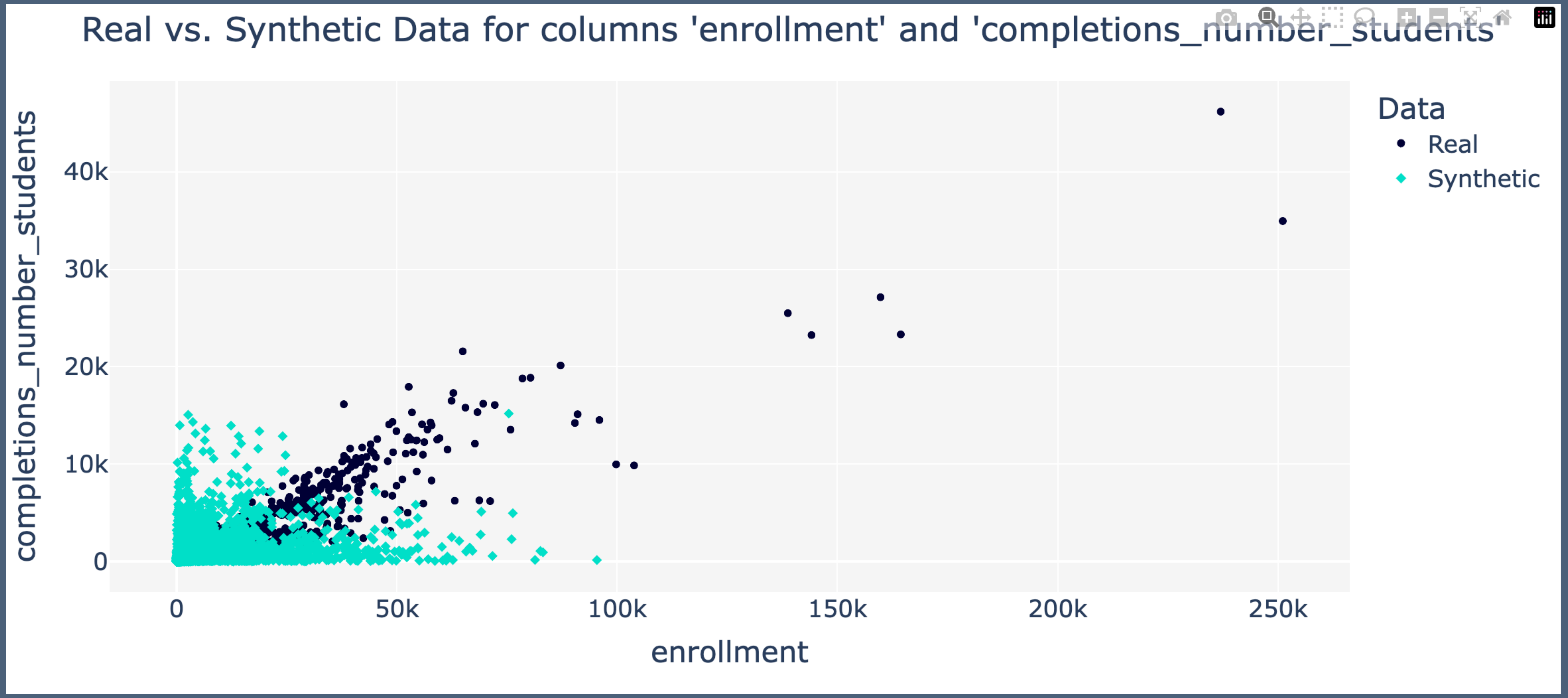
Metric	Training Set	Target	Value
MSE	Real	Enrollment	101,200,325.55
MSE	Synthetic	Enrollment	111,006,681.53
MSE	Combined	Enrollment	101,978,162.27
R^2	Real	Enrollment	0.19
R^2	Synthetic	Enrollment	0.11
R^2	Combined	Enrollment	0.19
MSE	Real	Completions	116,066,105.90
MSE	Synthetic	Completions	133,327,873.83
MSE	Combined	Completions	129,922,536.22
R^2	Real	Completions	-0.01
R^2	Synthetic	Completions	-0.07
R^2	Combined	Completions	-0.04

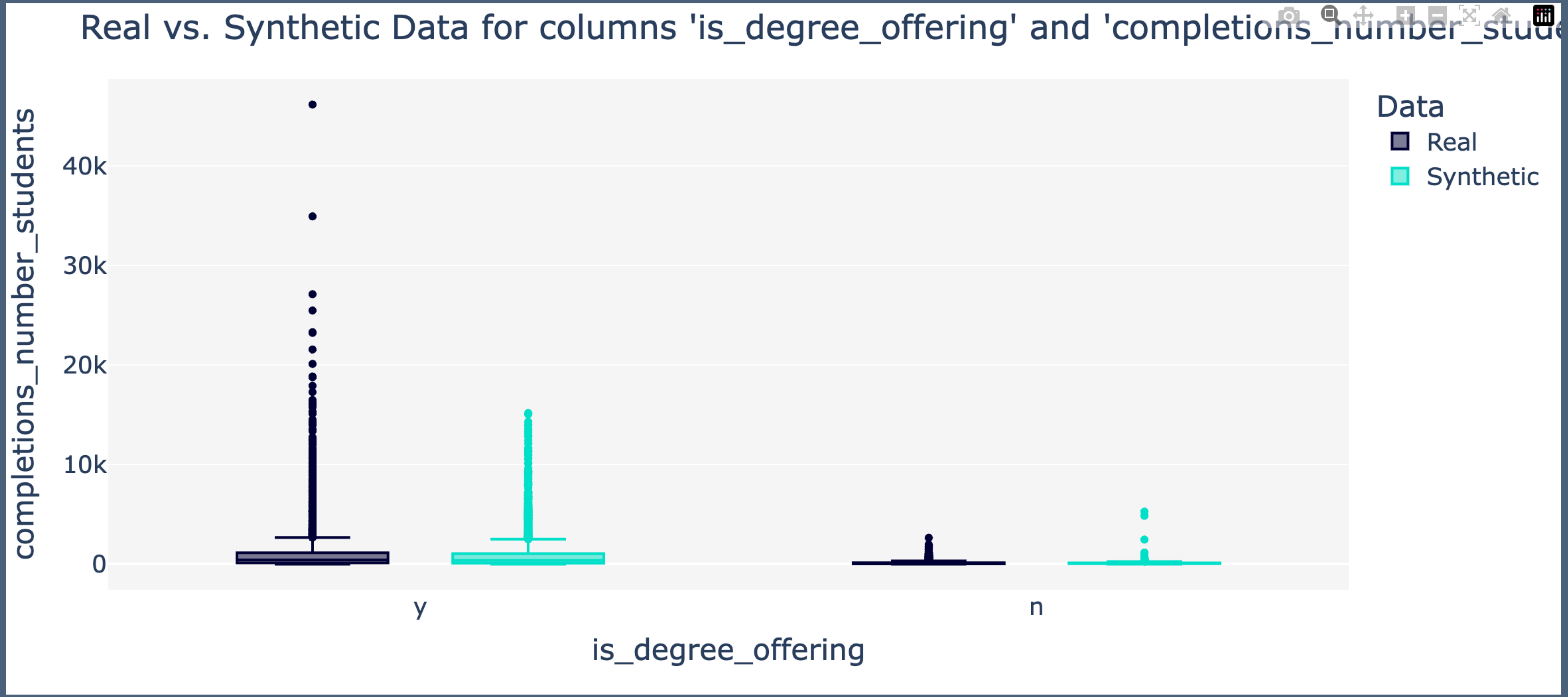
Evaluation Output





Evaluation Output





Timeline and Scope

- Timeline of 3-4 weeks to manage scope
- Focus on creating synthetic IPEDS data and exploring its quality and suitability for handling reductions in real IPEDS data
- Picking 3 domains of IPEDS data and further narrowing that down
- Database and scripts instead of more powerful software
- Adjusting end of pipeline to focus on exploring and evaluating generative modeling

Lessons and Questions

- The biggest challenge has been the complexity of the IPEDS data
- Where to put generative modeling in the pipeline?
- How to interpret generative modeling evaluation?
- How to incorporate student-level data?

Future Work

- More domains of IPEDS data
- Exploring generative modeling at different stages of pipeline
- More thorough evaluation process for generative modeling, including tracking history
- Utility functionality to facilitate pipeline expansion
- Incorporation of dedicated orchestration software or custom orchestration design