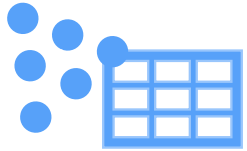
Generierung Synthetischer Daten mit dem SAS Data Maker

Gerhard Svolba

What is synthetic data?

Synthetic data is generated by applying a sampling technique to real-world data or by creating simulation scenarios where models and processes interact to create completely new data not directly taken from the real world.

– Gartner



Synthetic data and generative AI

Synthetic data has become important with the emergence of GenAI given GenAI models require large amounts of data for training.

Synthetic data generation can be considered a subset of GenAI given it generates data.

Synthetic data is also considered to be a privacy preservation enabler.

Organization: Nationwide Building Society
Industry: Financial Services
Region: UK

Synthetic data helped improve AI model accuracy, potentially reducing losses.

PROBLEM

- Machine learning credit scoring models must guide personal loan decisions in a fair, responsible manner
- Previous models using only real-life data faced limitations in accuracy, scalability, and data sensitivity

ACTION

- Tested synthetic data generated with SAS to improve machine learning model, comparing model accuracy versus models using only real-life data

RESULT

- 28% improvement in model accuracy with facilitated machine learning efforts, supporting potential reduction in losses

Organization: US Health Care Provider
Industry: Health Care
Region: United States

Synthetic data unlocks value from sensitive data without compromising privacy and security

PROBLEM

- Patient healthcare data is protected by privacy regulations, restricting access to data and increasing costs.
- These costs made performing analytics with the data cost-prohibitive.
- Provider must support an ecosystem of shared research and innovation involving universities and research groups.

ACTION

- Provide secure and unified platform which enables synthetic data generation to simulate patient behavior and outcomes, test treatment plans and choose optimal care paths.
- Platform also offers critical supporting capabilities such as data transformation, generation and post-processing.

RESULT

- Enables patients to benefit from data-driven research on synthetic clinical datasets
- Reduced risk to customer privacy

Industry: Banking
Region: Europe

Synthetic data enhanced customer privacy and collaboration with external vendors

PROBLEM

- Lack of safe and realistic test data. Privacy leakage concerns.
- Needed a way to share data with external vendors without risks to customer privacy and compliance.

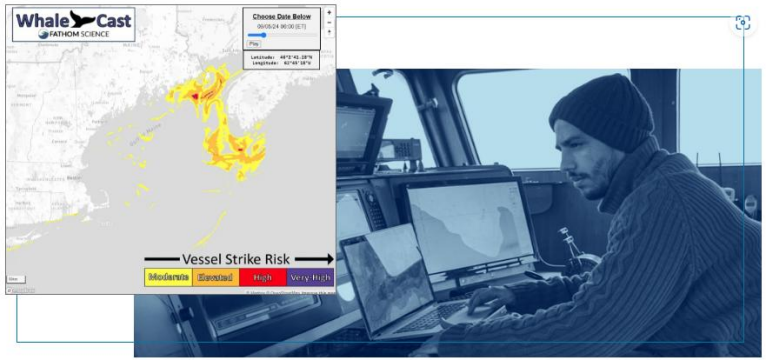
ACTION

- Generated synthetic data repository for on-demand test data.
- New products were generated in a sandbox with synthetic data.

RESULT

- Sandbox was safely accessed by over 700 external developers.
- Enhanced privacy, reduced risk, improved collaboration with external vendors.

Organization: Fathom Science
Industry: Marine Conservation
Region: United States



Synthetic data augmented insufficient real data, helping conserve right whales

PROBLEM

- North Atlantic right whales are endangered, partly due to being struck accidentally by boats
- Need a predictive heat map of whale movement to inform ship captains
- Insufficient whale sighting data to validate predictive model of whale movement

ACTION

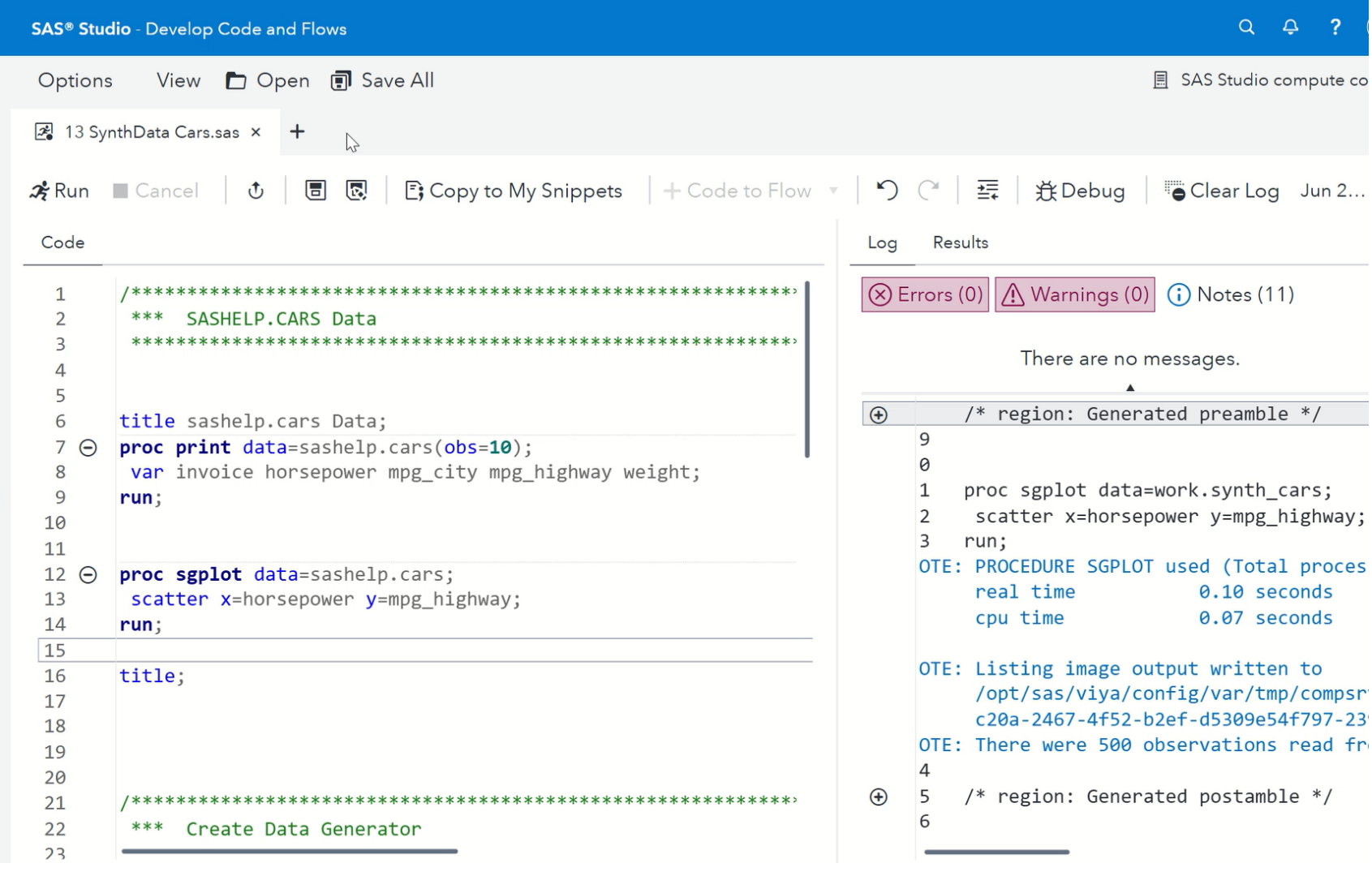
- Used SAS to generate synthetic data to achieve sufficient data for validation (expanded set from 40K to 500K data points).
- Fathom was able to validate their predictive model, increasing confidence using machine learning.

RESULT

- WhaleCast tool can be integrated into existing boat on-board touch screens, helping mariners better understand areas with greater risk of striking whales

SAS Viya Software Demo:

Erzeugung Synthetischer Daten mit PROC SIMNORMAL


SAS® Studio - Develop Code and Flows
Options
View
Open
Save All
SAS Studio compute co
13 SynthData Cars.sas
Run
Cancel
Copy to My Snippets
Code to Flow
Debug
Clear Log
Jun 2...
Code
1 /*****
2 *** SASHELP.CARS Data
3 *****
4
5
6 title sashelp.cars Data;
7 proc print data=sashelp.cars(obs=10);
8 var invoice horsepower mpg_city mpg_highway weight;
9 run;
10
11
12 proc sgplot data=sashelp.cars;
13 scatter x=horsepower y=mpg_highway;
14 run;
15
16 title;
17
18
19
20
21 /*****
22 *** Create Data Generator
23
Log
Results
Errors (0)
Warnings (0)
Notes (11)
There are no messages.
/* region: Generated preamble */
9
0
1 proc sgplot data=work.synth_cars;
2 scatter x=horsepower y=mpg_highway;
3 run;
OTE: PROCEDURE SGPLOT used (Total proces
real time 0.10 seconds
cpu time 0.07 seconds
OTE: Listing image output written to
/opt/sas/viya/config/var/tmp/compsr
c20a-2467-4f52-b2ef-d5309e54f797-23
OTE: There were 500 observations read fr
4
5 /* region: Generated postamble */
6

SAS Viya Software Demo:

Erzeugung Synthetischer Daten mit dem SAS DATA MAKER

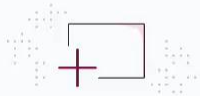
SAS® Data Maker

Your gateway to seamless synthetic data generation

New project

Projects

Search



Create a project to get started.

New project

How to

Overview

See an overview of SAS Data Maker.

Watch video

Get Started

Learn about SAS Data Maker.

View documentation

Provide Feedback

We would love to hear about your experience.

Submit feedback

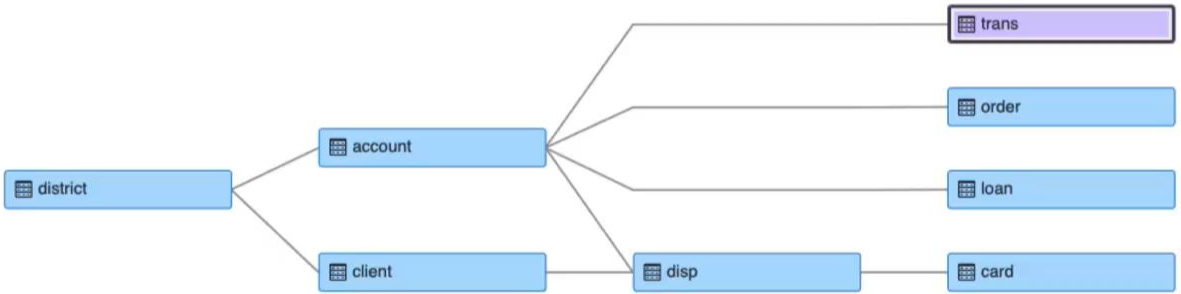
INNOVATE DEMO

Source Data | Training | Evaluation | Generate

Default model

- account
- card
- client
- disp
- district
- loan
- order
- trans

Structure | Columns



trans

Source:
banking/trans.csv.gz

Table type
Sequential

Sequential id
account_id

Sort by

Primary key
trans_id

Foreign keys

Key:	Target:
accoun...	ACCOUNT.acc...

+ Add foreign key

Date modified:
Dec 31, 1, 11:23 PM

Date created:
Apr 17, 2025, 5:20 PM

Size:
38.15 MB

INNOVATE DEMO

Source Data | Training | Evaluation | Generate

- All Tables
- account
- card
- client
- disp
- district
- loan
- order
- trans

Structure | Columns

Table	Column	Semantic Type
account	account_id	ID
account	district_id	Foreign key
account	frequency	Category
account	date	Date time
card	card_id	ID
card	disp_id	Foreign key
card	type	Category
card	issued	Date time

INNOVATE DEMO

Source Data | Training | Evaluation | Generate

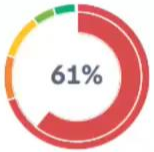
25 Apr 2025 at 14:55

This model has been trained. Make a copy to train with different settings. Make a Copy and Edit

Cross-table mutual information

Degree distribution

Density disclosure



Marginal distribution

Mutual information

Presence disclosure

Assessment Metrics | Generated Data Sample

Marginal distribution

Column account.frequency (99.5%)