Generating Valuable Dummy Data for Analytical Method Development

Tadashi Matsuno, Yoshitake Kitanishi, Shionogi & Co., Ltd.

SHIONOGI

Things to keep in mind

Here we focus only on the technical aspects and how to dummy data.

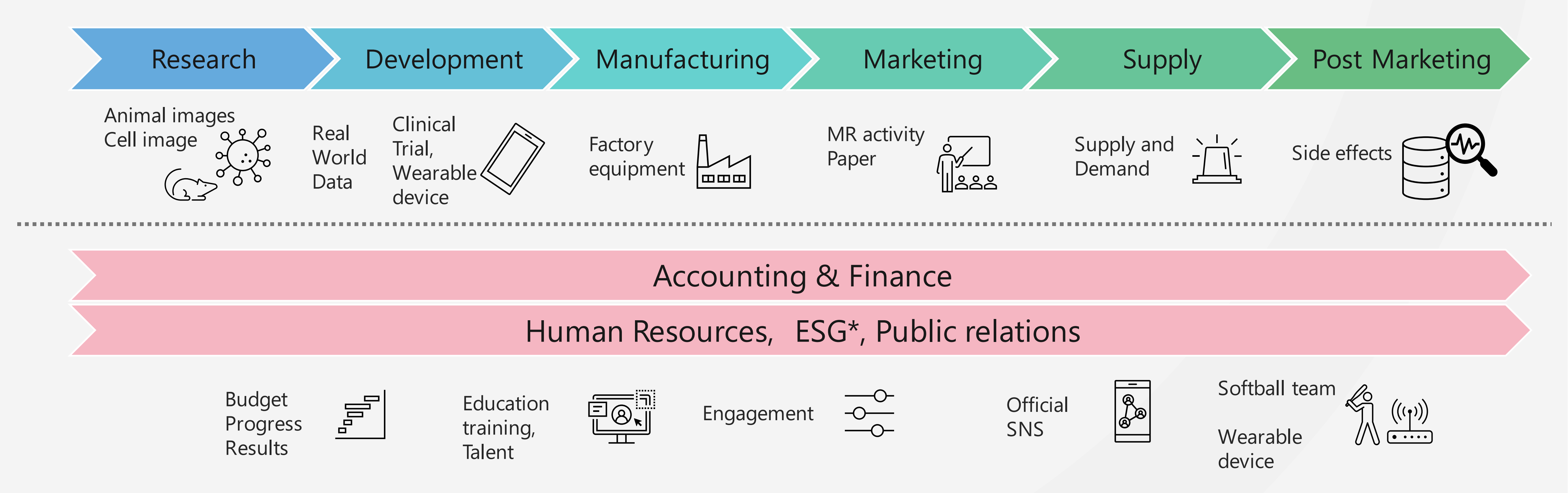
When actually using data, it is necessary to consider multiple aspects in addition to technology, such as laws and regulations, contracts, and an understanding of data providers.

Summary

- Background for considering methods for dummying data
- Scenarios to consider
- Exploration of methods for dummying one-record data
- Dummying data and evaluation
- About the future

Data related to SHIONOGI Data Science Department

The Data Science Department is promoting the integration of valuable and diverse internal data, as well as the development of data utilization infrastructure, advanced analysis techniques, and talent cultivation, in order to enable cross-functional utilization of these data in the value chain..



Research

Animal images
Cell image

Development

Real World Data

Clinical Trial, Wearable device

Manufacturing

Factory equipment

Marketing

MR activity
Paper

Supply

Supply and Demand

Post Marketing

Side effects

Accounting & Finance

Human Resources, ESG*, Public relations

Budget
Progress
Results

Education training, Talent

Engagement

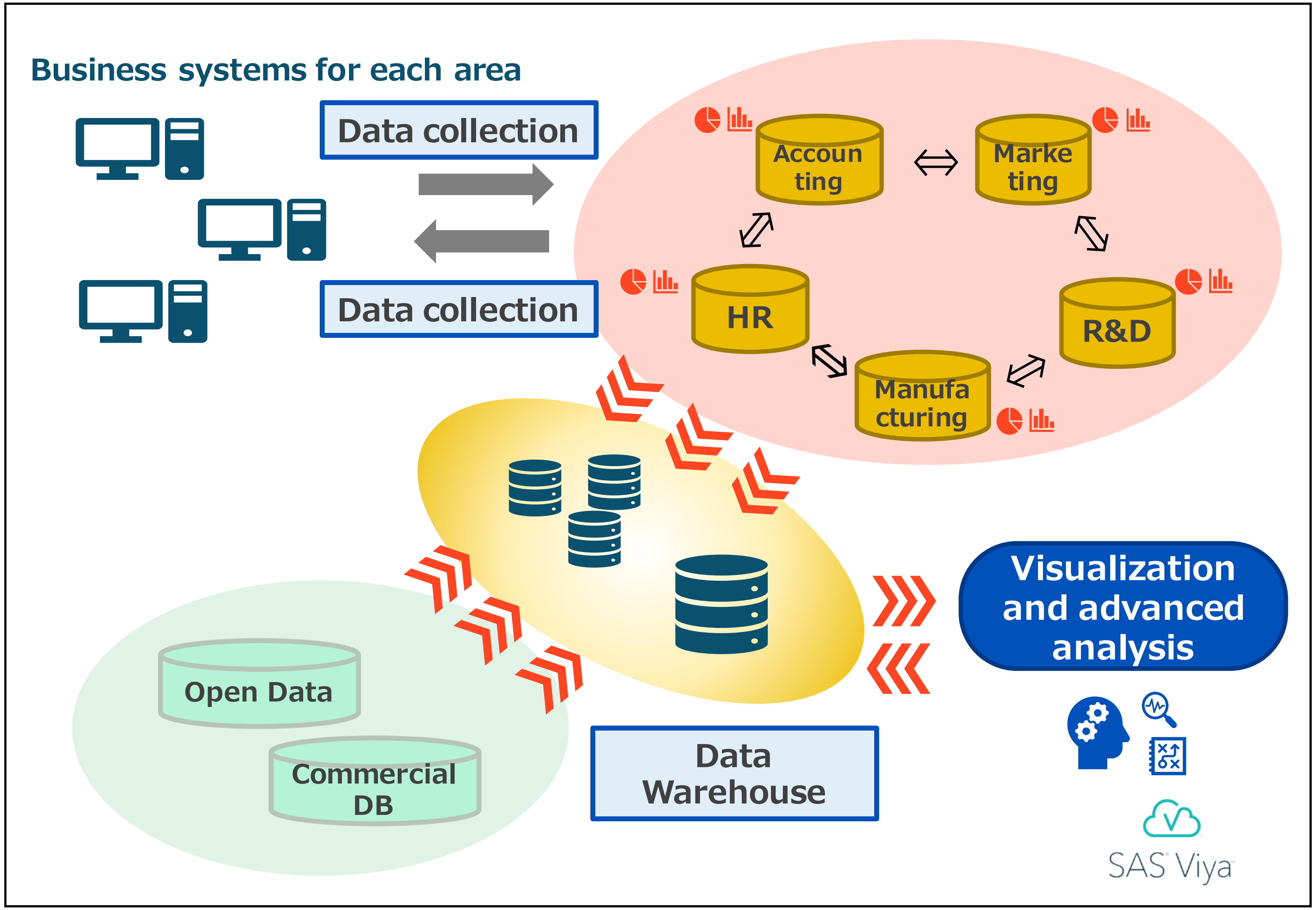
Official SNS

Softball team

Wearable device

*Environment, Social, Governance)

SHIONOGI Central Data Management Concept



Hub for data collection and distribution between business systems

Data warehouse that prepares data in a format that is easy to use.

- Data modernization
- Data cleansing
- Mapping conversion, integration
- Confidential management

Motivation

- Dummy data without using anonymous processing technology

Deliverables

✓It can be expected the same analysis as the original data without reducing the amount of information.
✓Unable to extract actual sensitive information (such as specific individuals)

Possibility of using dummy data

Utilization of data in a wide range of departments leading to decision-making

adding

Data analysis training material topics

Data for checking system operation

Concrete application

Maybe…

Data sharing with society

Participating in a data-driven society

Dummy data generation scenario

The dummy coding techniques and evaluation criteria for data sets used in analysis vary depending on their shape and purpose, and it is necessary to conduct verification for each of them.

- One Record（ e.g. Personnel information, subject background） **Today's Focus**
- Time series data (e.g. stock price, blood pressure for each subject)
- Relational data (e.g. RWD, SDTM, ADaM)
- Inflating data (e.g., improving machine learning modeling accuracy)

Scenario
One Record data dummyization

SHIONOGI

One record data

Consists of a row containing a unique identifier and associated information fields and attributes 2D data table

For Example…

product list	**PRODUCT ID**	**NAME**	**CATEGORY**	**PRICE**	**WEIGHT**
Employee information (1 person Record)	**EMP NO.**	**NAME**	**DEPARTMENT**	**POSITION**	**START DATE**
Subject background (DM)	**USUBJID**	**AGE**	**SEX**	**RACE**	**ARM**

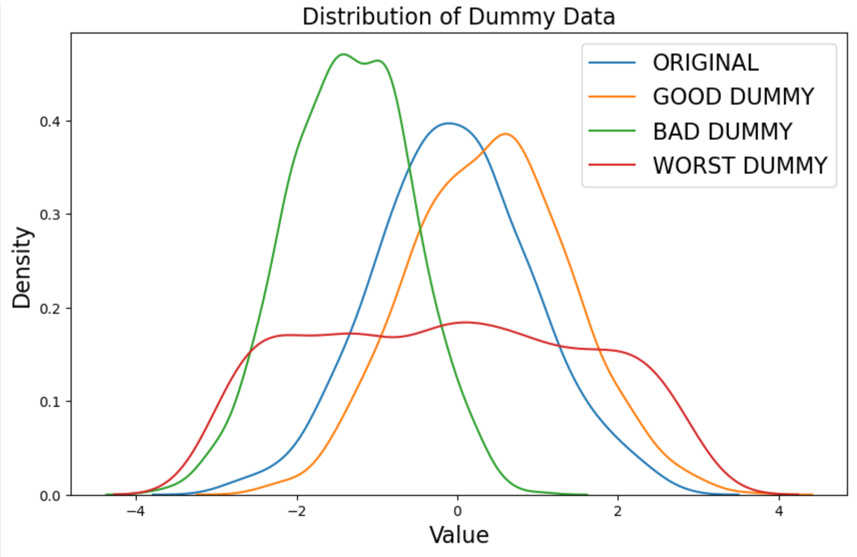
Analysis scene

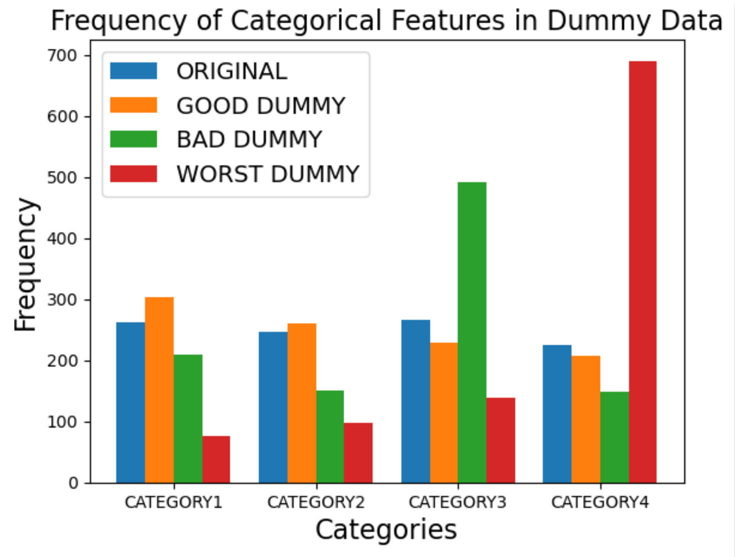
- ✓ Understand the characteristics of the entire data through summary statistics, distribution visualization, etc.
- ✓ Qualify data by joining with other tables using identifiers as keys

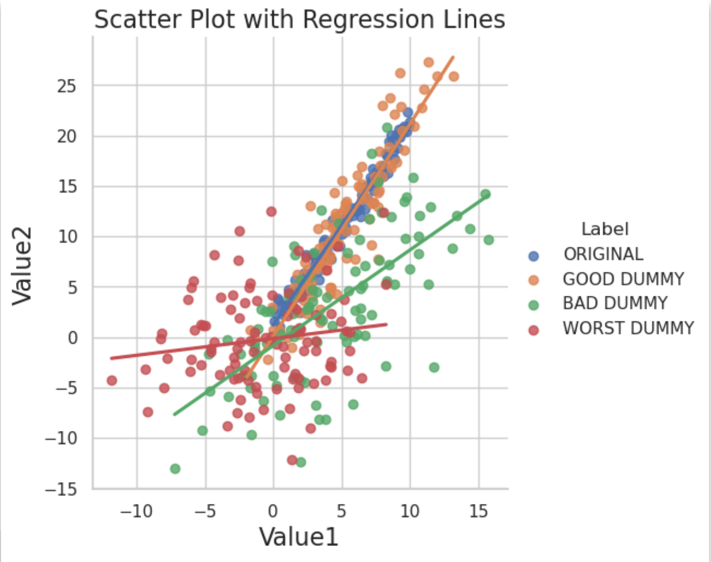
Dummyization of one-record data: What can be imitated?

- ✓ Distribution of numerical data
- ✓ Distribution of categorical values
- ✓ Relationship between features

⇒ These have the same characteristics as the original data, and the individual rows do not match.


Distribution of Dummy Data


Frequency of Categorical Features in Dummy Data


Scatter Plot with Regression Lines

Creating a sample dataset for the experiment

Created 600 lines of DM-like sample data for safe and secure verification.

*Age, height, and weight are not categorized and treated as numerical data
(for verification without using anonymous processing technology)

Feature value

- **ID (unique)**
- **Height:** Continuous value correlated with weight normal distribution
- **Weight:** Continuous value that correlates with height, only 40 kg or more
- **Age:** Discrete value, only over 20 years old
- **Sex:** Category (nominal scale) Male to female 1:1
- **Race/ethnicity:** correlated categories (nominal scale) imbalance
- **Drug assignment:** Category (nominal scale) 1:1 Uncorrelated
- **Test result 1:** Category (nominal scale) binary
- **Test result 2:** Category (nominal scale) binary

Sample "DM like" Dataset

ID	AGE	HEIGHT	WEIGHT	SEX	RACE	ETHNIC	ARM	TEST01	TEST02
SAMPLE0000	31	165.9	54.8	F	RACE1	ETHNIC2	PLACEBO	NEGATIVE	POSITIVE
SAMPLE0001	53	153.8	74.9	F	RACE3	ETHNIC1	PLACEBO	NEGATIVE	NEGATIVE
SAMPLE0002	57	158.0	66.4	F	RACE1	ETHNIC2	PLACEBO	NEGATIVE	POSITIVE
SAMPLE0003	42	173.1	77.6	F	RACE3	ETHNIC2	PLACEBO	NEGATIVE	NEGATIVE
SAMPLE0004	32	157.2	49.0	F	RACE1	ETHNIC2	PLACEBO	NEGATIVE	POSITIVE
SAMPLE0005	55	179.4	69.1	M	RACE1	ETHNIC2	PLACEBO	POSITIVE	POSITIVE
SAMPLE0006	65	167.1	69.4	M	RACE1	ETHNIC2	PLACEBO	POSITIVE	POSITIVE
SAMPLE0007	66	154.2	70.0	F	RACE1	ETHNIC2	PLACEBO	NEGATIVE	POSITIVE
SAMPLE0008	33	160.0	74.1	F	RACE1	ETHNIC2	PLACEBO	NEGATIVE	POSITIVE
SAMPLE0009	40	182.7	102.5	M	RACE3	ETHNIC1	PLACEBO	NEGATIVE	NEGATIVE
SAMPLE0010	44	169.8	48.6	F	RACE1	ETHNIC2	PLACEBO	NEGATIVE	POSITIVE

600 Rows （ACTIVE300: PLACEBO300）

How to make a one-record data table dummy

In order to perform dummyization while preserving the original characteristics, we tried the following method.

Missing imputation dummyization method

After intentionally generating missing values, kNN is used to complete missing values and process the data into data that has similar characteristics to the original data but is not completely identical.

Deep learning (GAN*)

Modeling the target data structure using two deep learning networks: a discriminator and a generator. This allows for the generation of synthetic data according to desired specifications.

LLM*(ChatGPT)

A language model built on a massive dataset. By conveying the desired features of the data to be generated through prompts (natural language instructions), the model produces output.

*Generative Adversarial Networks

*Large Language Models

Missing imputation dummyization method

A method of dummying data by intentionally missing and supplementing part of the data

①Generate missing values with an arbitrarily specified probability for each column

②Predict and impute missing values based on non-missing rows. kNN was used for prediction

SAMPLE

AGE	SEX	RACE	ETHNIC
21	F	RACE1	ETHNIC2
24	F	RACE3	ETHNIC1
35	F	RACE1	ETHNIC2
52	F	RACE3	ETHNIC2

Data spoiling p=0.2

AGE	SEX	RACE	ETHNIC
21	F	RACE1	ETHNIC2
24	F	RACE3	ETHNIC1
35	F	RACE1	ETHNIC2
	F	RACE3	ETHNIC2

AGE	SEX	RACE	ETHNIC
21	F	RACE1	ETHNIC2
24	F	RACE3	ETHNIC1
35	F	RACE1	ETHNIC2

AGE	SEX	RACE	ETHNIC
	F	RACE3	ETHNIC2

Learning Dataset

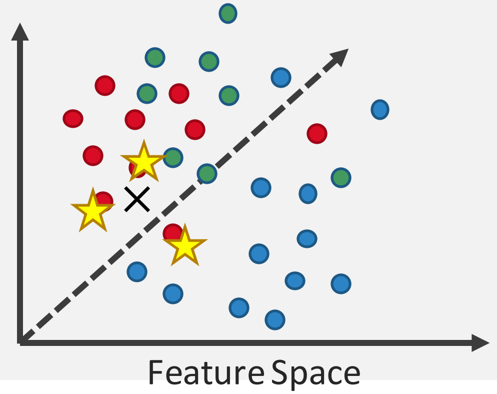
Target

AGE
21
24
35

Features

SEX	RACE	ETHNIC
F	RACE1	ETHNIC2
F	RACE3	ETHNIC1
F	RACE1	ETHNIC2

kNN



SEX	RACE	ETHNIC
F	WHITE	NOT HISPANIC

Estimate

AGE
33

Next Column…

AGE	SEX	RACE	ETHNIC
21	F	RACE1	ETHNIC2
24	F	RACE3	ETHNIC1
35	F	RACE1	ETHNIC2
33	F	RACE3	ETHNIC2

Expectations and concerns about this method

- Applicable to both classification and regression
- Tends to approach the average value/mode
- Original data may not be processed

GAN（Generative Adversarial Networks）

How to model table data using generative adversarial networks

CTGAN*: An algorithm that applies GAN, which has been mainly used for image generation, to table data

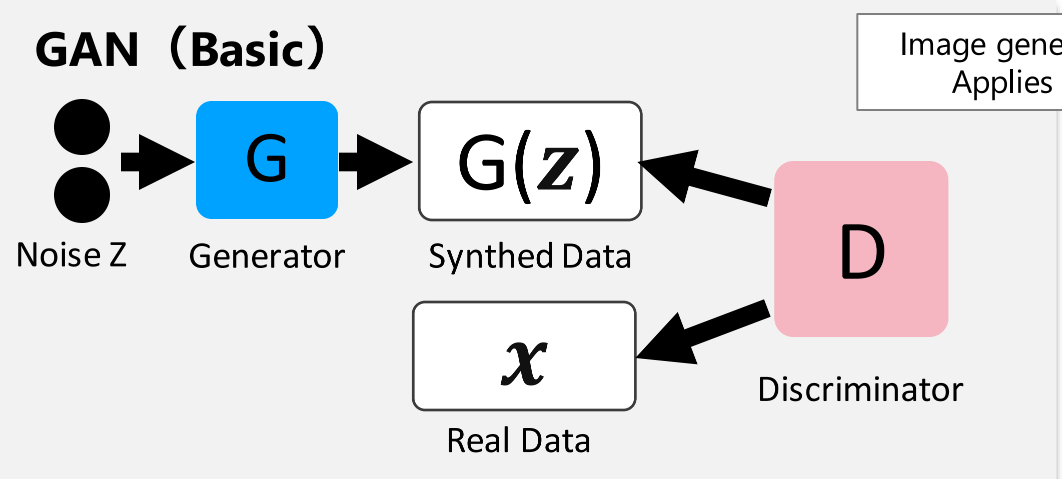
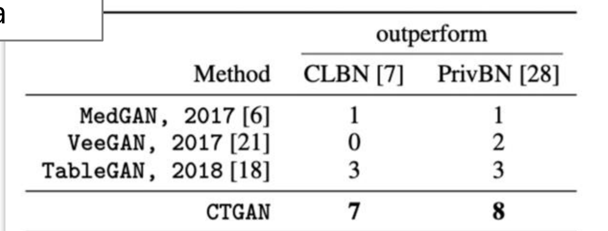


Image generation algorithm
Applies to table data



	outperform	
Method	CLBN [7]	PrivBN [28]
MedGAN, 2017 [6]	1	1
VeeGAN, 2017 [21]	0	2
TableGAN, 2018 [18]	3	3
CTGAN	**7**	**8**

Outperformed conventional methods for table data generation

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling Tabular data using Conditional GAN. arXiv preprint arXiv:1907.00503.

Expectations and concerns about this method

- You can model the distribution of table data and complex relationships.
- Less risk of data being the same as the original
- Very large computational resources
- Difficult to interpret internal calculations of the network

*CTGAN: Conditional Tabular Generative Adversarial Networks

Large language model（ChatGPT）

- In a secure OpenAI environment, explain the data set using prompts and output it as SAS program code.
- Add correction instructions after checking the distribution of generated data

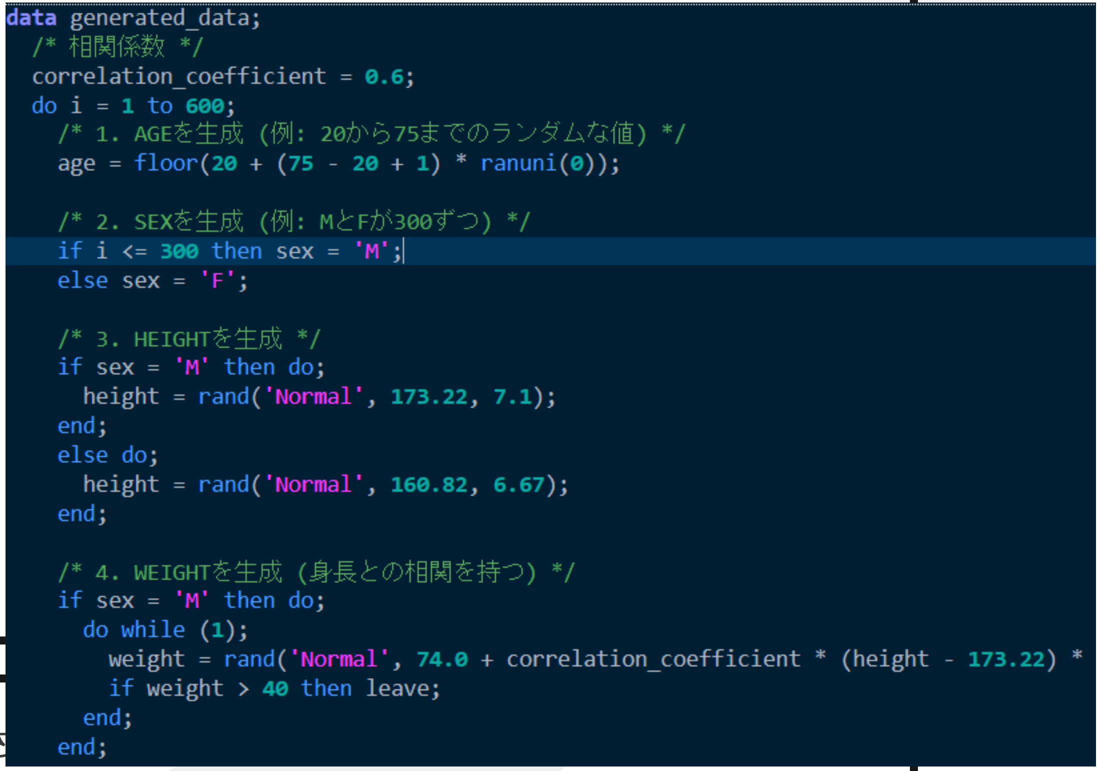
次の特徴をもつダミーデータを生成するSASコードを生成してください

カラムは、AGE,HEIGHT,WEIGHT,SEX,ARM,RACE,ETHNICです。
600rowあります。
AGEは20以上の整数です。
ARMはACTIVEとPLACEBOの2値で300ずつです。他のカラムと相関がありません。
SEXはMとFが300ずつです。
HEIGHTは平均173.22、標準偏差7.1・・・・

以下略

Modifying instruction

HEIGHTとWEIGHTの相関がないようです。相関係数0.6を想定して修正してくださ



```
data generated_data;
  /* 相関係数 */
  correlation_coefficient = 0.6;
  do i = 1 to 600;
    /* 1. AGEを生成 (例: 20から75までのランダムな値) */
    age = floor(20 + (75 - 20 + 1) * ranuni(0));

    /* 2. SEXを生成 (例: MとFが300ずつ) */
    if i <= 300 then sex = 'M';
    else sex = 'F';

    /* 3. HEIGHTを生成 */
    if sex = 'M' then do;
      height = rand('Normal', 173.22, 7.1);
    end;
    else do;
      height = rand('Normal', 160.82, 6.67);
    end;

    /* 4. WEIGHTを生成 (身長との相関を持つ) */
    if sex = 'M' then do;
      do while (1);
        weight = rand('Normal', 74.0 + correlation_coefficient * (height - 173.22) *
        if weight > 40 then leave;
      end;
    end;
```

Generated data table

Using the method described above, we generated dummy data of 600 rows each.

Drug allocation: active drug: 300 patients each with placebo

compared to sample data

make an evaluation

ORIGINAL

AGE	HEIGHT	WEIGHT	SEX	RACE	ETHNIC	ARM
31	165.9	54.8	F	RACE1	ETHNIC2	PLACEBO
53	153.8	74.9	F	RACE3	ETHNIC1	PLACEBO
57	158	66.4	F	RACE1	ETHNIC2	PLACEBO
42	173.1	77.6	F	RACE3	ETHNIC2	PLACEBO
32	157.2	49	F	RACE1	ETHNIC2	PLACEBO
55	179.4	69.1	M	RACE1	ETHNIC2	PLACEBO
65	167.1	69.4	M	RACE1	ETHNIC2	PLACEBO
66	154.2	70	F	RACE1	ETHNIC2	PLACEBO
33	160	74.1	F	RACE1	ETHNIC2	PLACEBO
40	182.7	102.5	M	RACE3	ETHNIC1	PLACEBO

IMPUTING

AGE	HEIGHT	WEIGHT	SEX	RACE	ETHNIC	ARM
31	165.9	57.9	M	RACE1	ETHNIC2	PLACEBO
53	163	74.9	F	RACE3	ETHNIC1	PLACEBO
47	158	62.5	F	RACE1	ETHNIC2	PLACEBO
53	166	77.6	F	RACE3	ETHNIC2	PLACEBO
32	157.2	49	F	RACE1	ETHNIC2	PLACEBO
55	171	69.1	M	RACE1	ETHNIC2	PLACEBO
65	167.1	69.4	M	RACE1	ETHNIC2	PLACEBO
66	154.2	70	F	RACE1	ETHNIC2	PLACEBO
46	160.3	64.2	F	RACE1	ETHNIC2	PLACEBO
40	182.7	102.5	M	RACE3	ETHNIC1	PLACEBO

GAN

AGE	HEIGHT	WEIGHT	SEX	RACE	ETHNIC	ARM
40	164	65.1	F	RACE1	ETHNIC2	PLACEBO
27	157	61.6	F	RACE1	ETHNIC2	PLACEBO
48	172.4	78.5	M	RACE1	ETHNIC2	PLACEBO
62	150.5	52.5	F	RACE1	ETHNIC2	PLACEBO
58	159.7	53.7	F	RACE1	ETHNIC2	PLACEBO
32	158.2	78	M	RACE4	ETHNIC1	PLACEBO
68	172.3	69.6	M	RACE1	ETHNIC2	PLACEBO
51	171.8	99	F	RACE2	ETHNIC2	PLACEBO
48	159.9	58.4	F	RACE3	ETHNIC1	PLACEBO
36	156.9	86.8	F	RACE4	ETHNIC1	PLACEBO

LLM

AGE	SEX	HEIGHT	WEIGHT	RACE	ETHNIC	ARM
20	M	168.3	58.4	RACE1	ETHNIC2	PLACEBO
36	M	181.5	80.7	RACE1	ETHNIC2	PLACEBO
32	M	175.2	73.2	RACE3	ETHNIC1	PLACEBO
28	M	170.5	85.3	RACE1	ETHNIC2	PLACEBO
36	M	173.2	51.0	RACE4	ETHNIC1	PLACEBO
33	M	179.0	69.7	RACE1	ETHNIC2	PLACEBO
59	M	168.6	71.1	RACE3	ETHNIC2	PLACEBO
49	M	176.3	86.4	RACE3	ETHNIC2	PLACEBO
21	M	180.3	80.9	RACE1	ETHNIC2	PLACEBO
30	M	179.2	68.8	RACE1	ETHNIC2	PLACEBO

Evaluation of dummy data

The generated dummy data was evaluated using the following indicators:

①Distribution similarity

1. Comparison by visualizing distribution
2. Similarity evaluation using JS divergence*
   - Supports continuous values and categorical values
   - Unlike a test, it can be evaluated relatively.

JS divergence*
An indicator that shows the difference between two probability distributions. It is KL divergence with symmetry

$$KL(P||Q) = \int P(x)(logP(x) - logQ(x))$$

$$JS(P||Q) = \frac{1}{2}(KL(P||R) + KL(Q||R))$$

$$R(X = x) = \frac{1}{2}(P(X = x) + Q(X = x))$$

②<u>Relationships between columns</u>

- Numerical value x Numerical value ⇒ Pearson's R
- Numeric value x category value (nominal) ⇒ Correlation ratio
- Category value x category value (nominal) ⇒ Cramer's V

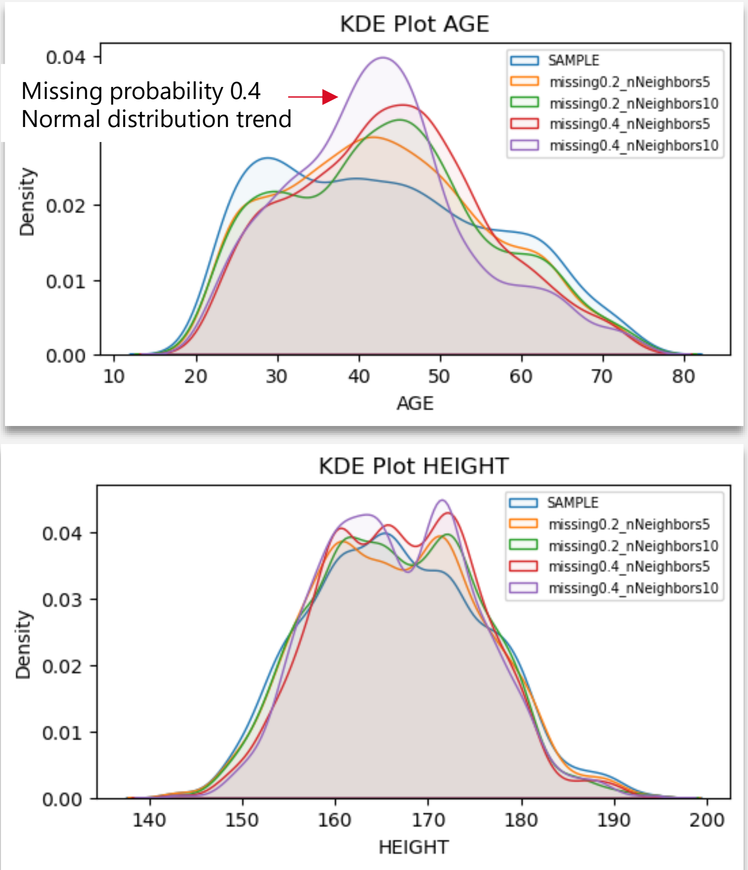
③Confidentiality risk

- Percentage of rows that match the original data in all columns

*Jensen-Shannon Divergence
*Kullback-Leibler Divergence

Distribution similarity ① Visualization of numerical columns (AGE, HEIGHT)

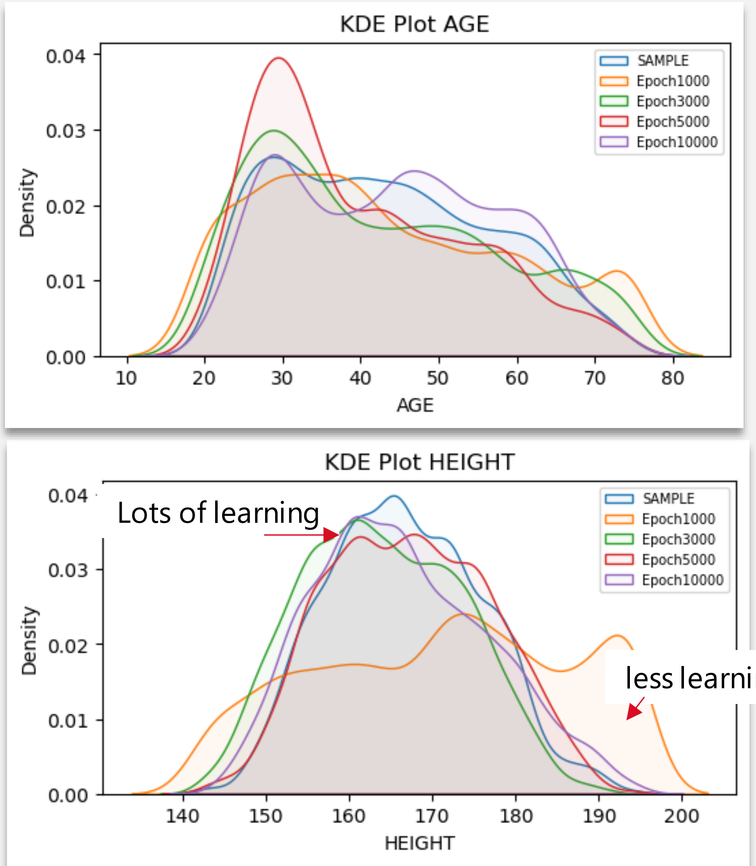
IMPUTING



Missing probability 0.4
Normal distribution trend

There is a tendency to converge to a normal distribution due to missing data imputation.

GAN

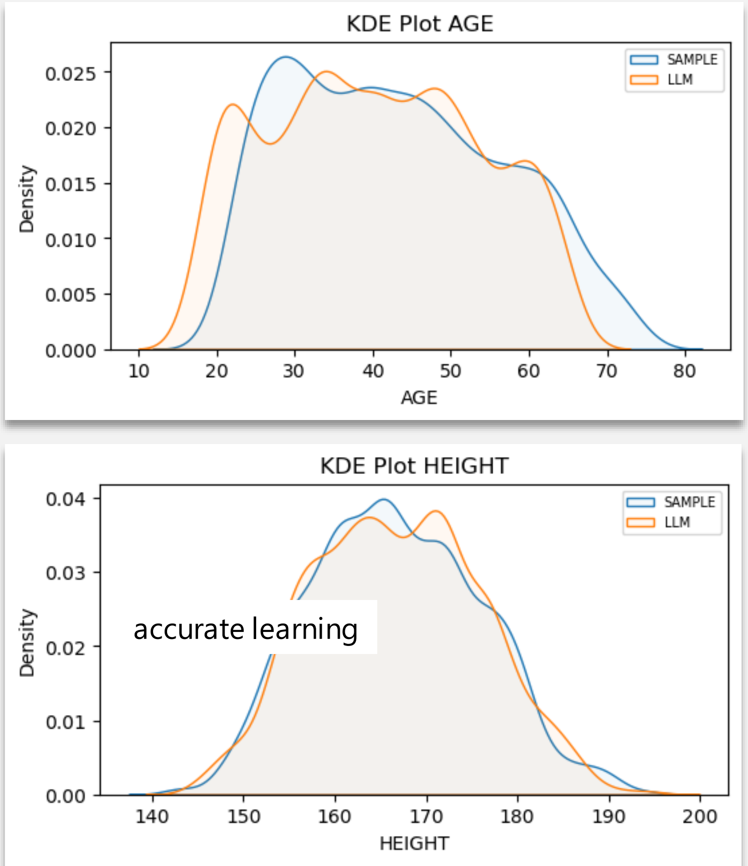


Lots of learning

less learning

The distribution will be captured accurately according to the number of learning times.

LLM

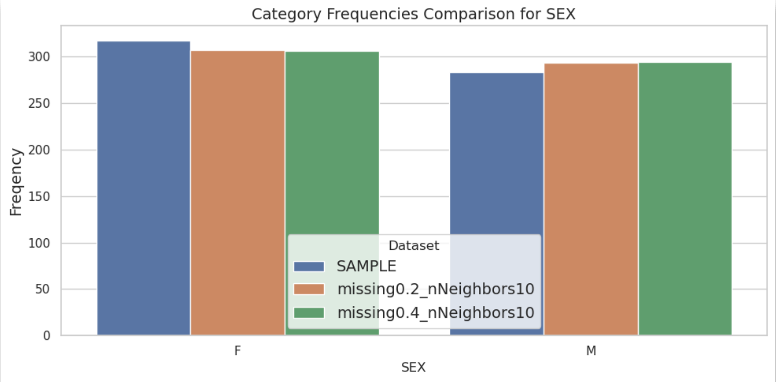


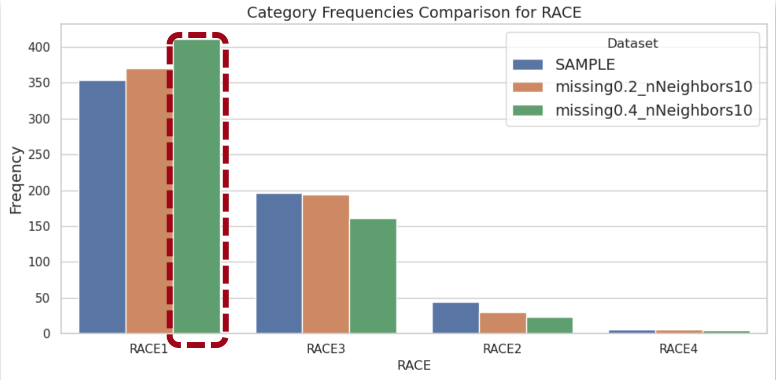
accurate learning

Reproduced the distribution relatively accurately

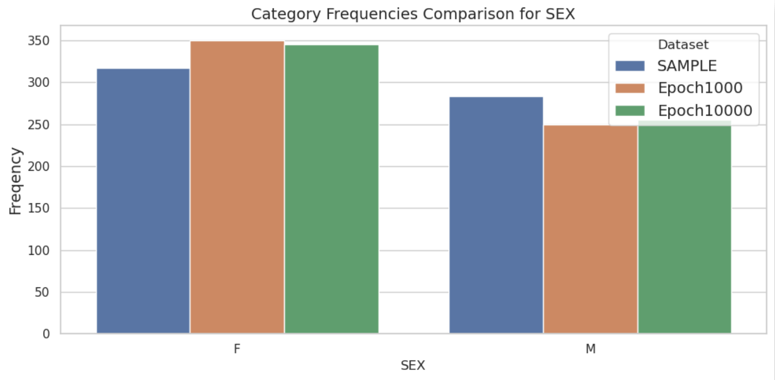
Distribution similarity ②
Visualization of category columns (nominal) (SEX, RACE)

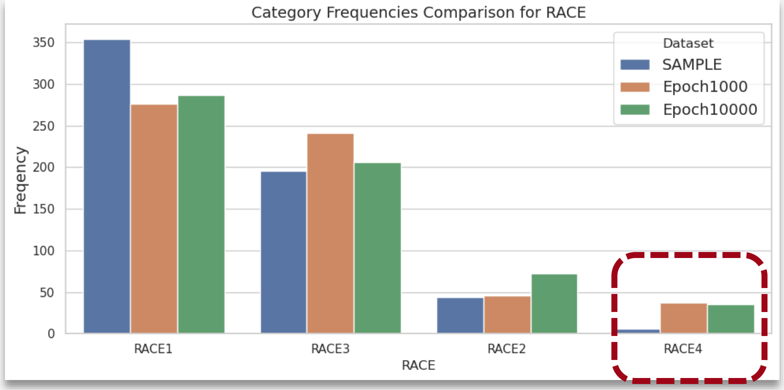
IMPUTING



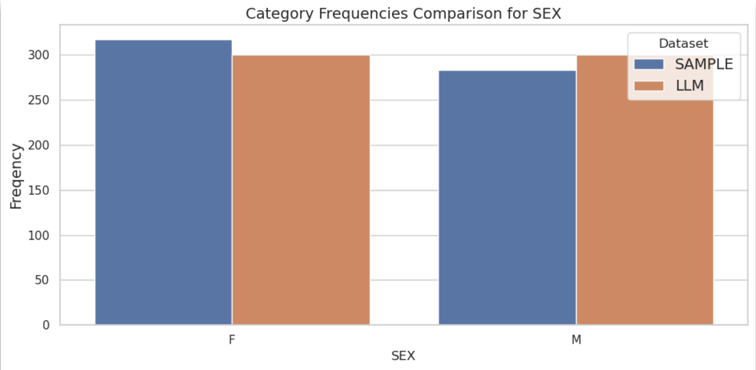


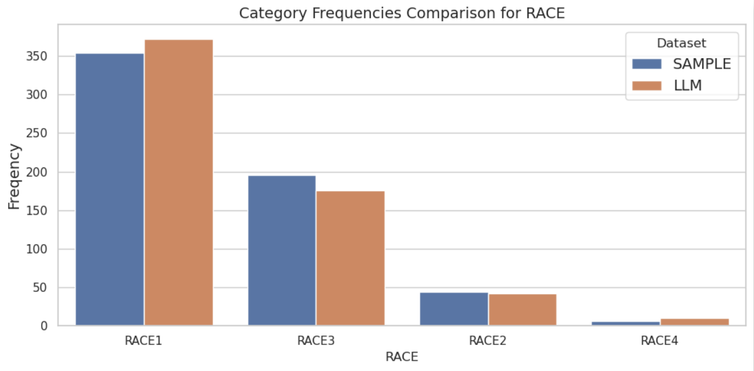
GAN





LLM





There is a tendency for majority selection to increase when imputing missing data.

There was a tendency to disproportionately increase the frequency of minorities.

Reproduced the distribution relatively accurately

Distribution similarity ③ Quantitative evaluation

We calculated the JS divergence value to quantitatively evaluate the similarity of the distributions.

It is close to 0 for all methods, and it is considered that the similarity with the original data is high.

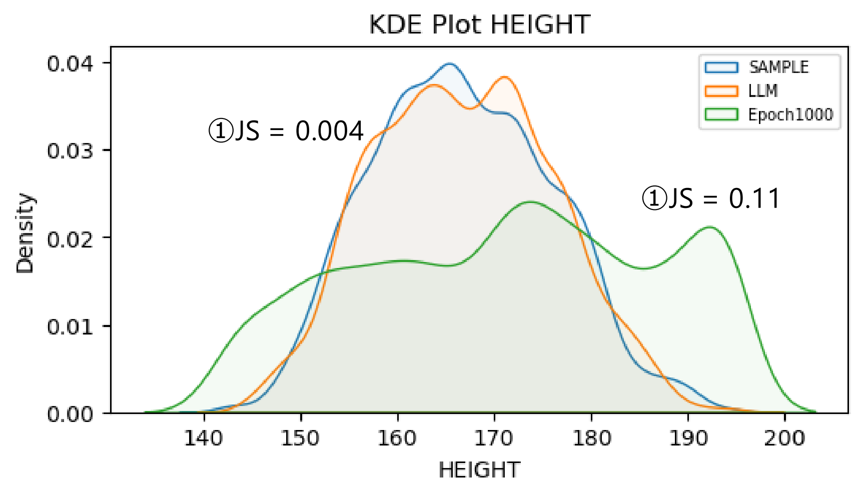
JS divergence value of each dummy data against sample data

Dummy DATA	Paremeter	Age	Height	Weight	Sex	Race	Ethnic
IMPUTING	Missing rate20%	0.0065	0.0052	0.0018	0.0001	0.0013	0.0000
IMPUTING	Missing rate40%	0.0230	0.0083	0.0067	0.0002	0.006	0.0007
GAN	Learning Epoch 1000	0.0354	① 0.11	0.0292	0.0015	0.016	0.0043
GAN	Learning Epoch 10000	0.0066	0.0066 ②	0.0152	0.0011	0.015	0.0050
LLM	-	0.027	0.004	0.01	0.01	0.001	0.0002

$$0 \leqq \text{JS} \leqq 0.69 (\fallingdotseq \log_e 2)$$

Perfect match completely different

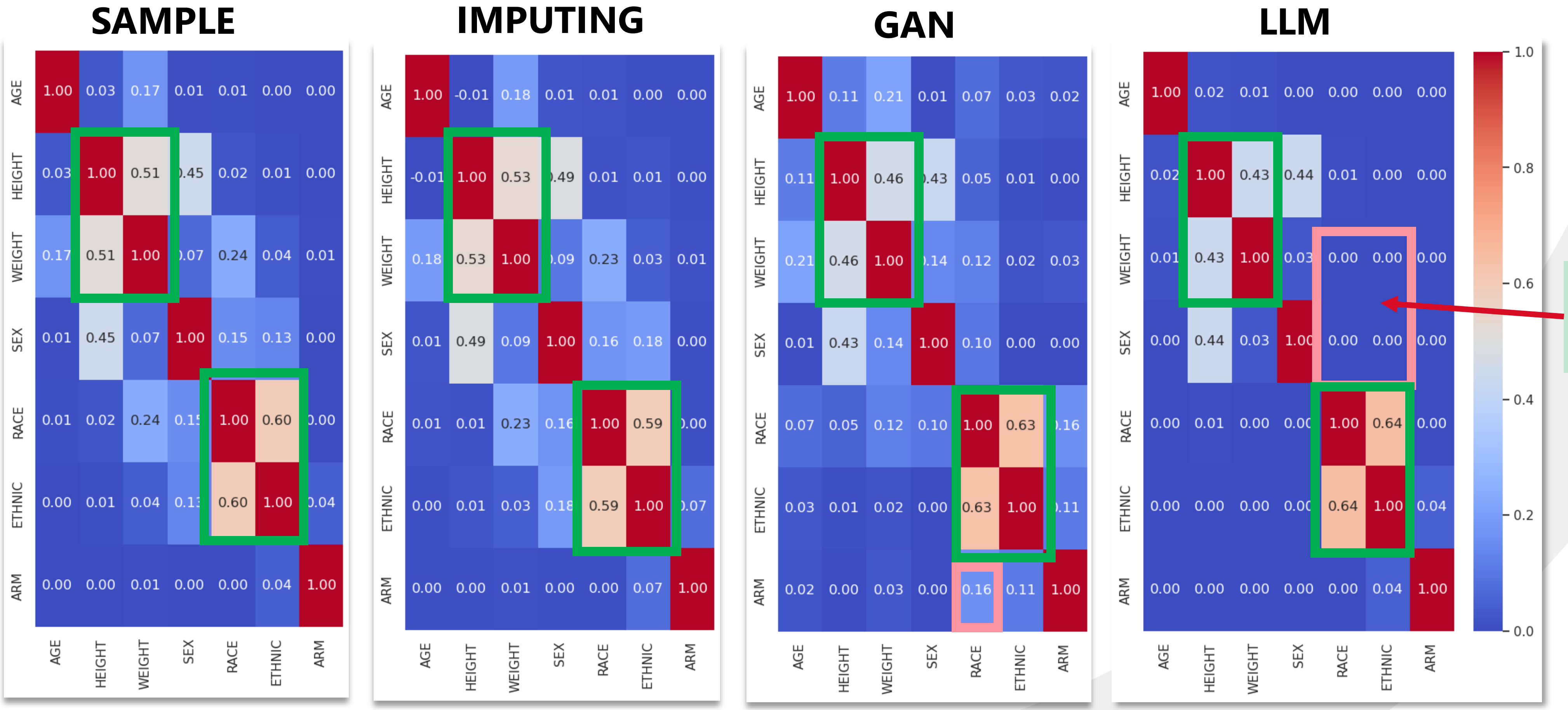
参考 JS value and actual distribution



Relationship between columns

In both cases, the relationships between columns generally showed the same tendency as the original.

- ✓ For GAN, there was a slight correlation with race in drug allocation.
- ✓ Elements not included in the prompt, such as the relationship between race and height, were not replicated in the LLM.



SAMPLE

	AGE	HEIGHT	WEIGHT	SEX	RACE	ETHNIC	ARM
AGE	1.00	0.03	0.17	0.01	0.01	0.00	0.00
HEIGHT	0.03	1.00	0.51	.45	0.02	0.01	0.00
WEIGHT	0.17	0.51	1.00	.07	0.24	0.04	0.01
SEX	0.01	0.45	0.07	1.00	0.15	0.13	0.00
RACE	0.01	0.02	0.24	0.15	1.00	0.60	.00
ETHNIC	0.00	0.01	0.04	0.13	0.60	1.00	.04
ARM	0.00	0.00	0.01	0.00	0.00	0.04	1.00

IMPUTING

	AGE	HEIGHT	WEIGHT	SEX	RACE	ETHNIC	ARM
AGE	1.00	-0.01	0.18	0.01	0.01	0.00	0.00
HEIGHT	-0.01	1.00	0.53	.49	0.01	0.01	0.00
WEIGHT	0.18	0.53	1.00	.09	0.23	0.03	0.01
SEX	0.01	0.49	0.09	1.00	0.16	0.18	0.00
RACE	0.01	0.01	0.23	0.16	1.00	0.59	.00
ETHNIC	0.00	0.01	0.03	0.18	0.59	1.00	.07
ARM	0.00	0.00	0.01	0.00	0.00	0.07	1.00

GAN

	AGE	HEIGHT	WEIGHT	SEX	RACE	ETHNIC	ARM
AGE	1.00	0.11	0.21	0.01	0.07	0.03	0.02
HEIGHT	0.11	1.00	0.46	.43	0.05	0.01	0.00
WEIGHT	0.21	0.46	1.00	.14	0.12	0.02	0.03
SEX	0.01	0.43	0.14	1.00	0.10	0.00	0.00
RACE	0.07	0.05	0.12	0.10	1.00	0.63	.16
ETHNIC	0.03	0.01	0.02	0.00	0.63	1.00	.11
ARM	0.02	0.00	0.03	0.00	0.16	0.11	1.00

LLM

	AGE	HEIGHT	WEIGHT	SEX	RACE	ETHNIC	ARM
AGE	1.00	0.02	0.01	0.00	0.00	0.00	0.00
HEIGHT	0.02	1.00	0.43	0.44	0.01	0.00	0.00
WEIGHT	0.01	0.43	1.00	0.03	0.00	0.00	0.00
SEX	0.00	0.44	0.03	1.00	0.00	0.00	0.00
RACE	0.00	0.01	0.00	0.00	1.00	0.64	0.00
ETHNIC	0.00	0.00	0.00	0.00	0.64	1.00	0.04
ARM	0.00	0.00	0.00	0.00	0.00	0.04	1.00

Race x height 0
Elements not prompted for

Pearson's R :
Continuous vs Continuous

Correlation Ratio :
Categorical vs Continuous

Cramer's V :
Categorical vs Categorical

Confidentiality risk assessment

Calculate the percentage of rows that completely match the original data
(also examine the effects of increasing or decreasing the number of columns)

- ✓ GAN,LLM was 0 regardless of the number of columns, showing excellent results.
- ✓ Although the missing data imputation method can be expected to be alleviated by adding columns, it was suggested that the risk is higher than other methods.

Percentage of rows that exactly match the sample data in all columns

DATA	Parameter	7columns	8columns	9columns
IMPUTING	Missing rate20%	44%	42%	40%
IMPUTING	Missing rate40%	21%	19%	17%
GAN	Learning Epoch 1000	0%	0%	0%
GAN	Learning Epoch 10000	0%	0%	0%
LLM	-	0%	0%	0%

7columns:
'AGE', 'HEIGHT', 'WEIGHT', 'SEX', 'RACE', 'ETHNIC','ARM'

8columns:
add "TEST1"

9columns:
add "TEST02"

TEST1	TEST2
NEGATIVE	POSITIVE
POSITIVE	NEGATIVE
NEGATIVE	POSITIVE
NEGATIVE	NEGATIVE

∗TEST:binary

One record data dummy evaluation

With either method, it was possible to create dummy data that captured the characteristics of the original data.
I would like to use them flexibly depending on the purpose, taking into consideration the characteristics of each.

Missing imputation method

- ✓ Easy to maintain relationships between columns, suitable for modeling purposes
- ✓ Stable results can be obtained regardless of data size
- ✓ Be especially careful when including personal information.

GAN

- ✓ It has low confidentiality risk and can be expected to be applied even when personal information is involved.
- ✓ Calculations are complex and output is unstable depending on the amount of data

LLM

- ✓ Easy to use due to low security risk and ease of calculation
- ✓ Only if the data can be expressed in prompts

Case	IMPUTING	GAN	LLM
Distribution of each column	Tendency to approach mean value/mode	Good	Good
Relationship between columns	Good	Imbalance data is partially inaccurate	Depends on prompt
Confidentiality risk	A certain number is the same as the original data	Good	Good

Use depending on usage/purpose

Summary

In order to promote the utilization of internal data, we are verifying data dummyization methods.

We were able to make One Record data into a dummy using the three methods discussed this time.
When applying to business, we want to select an appropriate method based on the intended use.
We will continue to explore and verify methods using a variety of data.

- One Record (e.g. personnel information, subject background)
- Time series data (e.g. stock price, blood pressure for each subject)
- Relational data (e.g. RWD, SDTM, ADaM)
- Inflating data (e.g., improving machine learning modeling accuracy, anomaly detection, etc.)

Future verification