AUGUST 2023

SYNTHETIC DATA: WHAT ARE THEY AND WHAT ARE THEY USED FOR?

GOBIERNO DE ESPAÑA
VICEPRESIDENCIA PRIMERA DEL GOBIERNO
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL
red.es
Oficina del Dato
Iniciativa aporta
datos.gob.es

Content

Content prepared by Alejandro Alija,

Expert in digital transformation and open data.

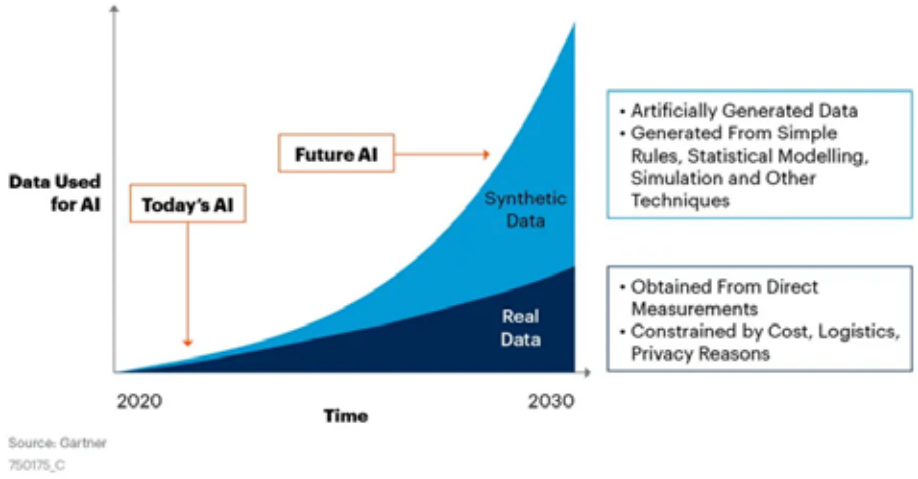
INTRODUCTION

Despite living in the age of data, contrary to what it may seem, **one of the main problems we face in building new digital products and services is the difficulty of accessing** valuable data in the context of that new product or service. Yes, strange as it may seem, there are many occasions when it is not easy to access quality data to understand a process or a system from a data perspective. Sometimes data exists but it is fragmented to such an extent that it is not possible to get a clear view of what is happening in an organisation, a product or a process. In other cases, we know that **the data exists**, but it is virtually **inaccessible for security or privacy reasons**. This is usually the case for **the most sensitive personal data such as medical or banking data**. From another angle, it is sometimes easy to find and even access the actual data in their respective sources, but it might be economically unfeasible or very expensive to use them. For example, we may have access to large databases of graphical resources such as images or videos with which to train AI algorithms, however, it is necessary to pay access costs and have expensive computing resources to import them into our system or training environment (extract them from their sources, store them, catalogue them, etc.). For these and many other reasons, **so-called synthetic data have been around for a long time**. So much so that Gartner, in an article published in the Wall Street Journal in 2021, claimed that by 2024, 60% of the data used for the development of artificial intelligence and analytics projects would be generated synthetically.

In this report, we try to **delve deeper into the field of synthetic data, explaining what it is and in what types of situations or use cases it is useful and necessary**. We will analyse **their benefits** compared to "real data" and **explain by means of a practical example how they are generated and a possible use of** them. Let's get started!

1. WHAT ARE SYNTHETIC DATA AND WHAT ARE THEY USEFUL FOR?

In the digital era in which we live, data generation and analysis have become fundamental to the **development of new technologies, products, and services**. One of the decisive factors why the COVID-19 vaccine could be achieved in such a short space of time had to do with the analysis of massive data from clinical trials, epidemiological data, mathematical simulation models, etc. This is something that had not been possible at any previous point in history. This is just one example, because, nowadays, any kind of important decision, in almost all organisations, is made on the basis of available data. However, as discussed in the introduction, in many cases, **accessing real data can be a challenge due to privacy restrictions, confidentiality or simply the unavailability of complete and up-to-date information.**



Original source: Gartner. From Forbes' "Synthetic Data Is About To Transform Artificial Intelligence"

In this context, **synthetic data has emerged as a promising solution. Synthetic data, unlike real data, is artificially fabricated information, rather than information generated by real-world events**. Synthetic data, among other uses, **are designed to mimic the characteristics and distributions of real data, without containing personal or sensitive information** that could identify individuals or compromise their privacy. This data **is created using algorithms and generation techniques** that **preserve the structure, relationships, and statistical properties** of the original data, **providing a safe and reliable alternative** for **analysis, experimentation**, and **training of** artificial intelligence **models**. Synthetic data is not a new idea . What is new is that they are now approaching a critical tipping point in terms of real-world impact. **It is about to change the whole value chain**, and the whole set of artificial intelligence technologies, a process that will have immense economic implications. Just look at the flurry of developments related to generative AI technologies in recent months.

Synthetic data is artificially fabricated information rather than information generated by real-world events. Synthetic data **are designed to mimic the characteristics and distributions of real data, without containing personal or sensitive information** that could identify individuals or compromise their privacy. These data **are created using algorithms and generation techniques** that **preserve the structure, relationships, and statistical properties** of the original data, **providing a safe and reliable alternative** for analysis, experimentation and training of Artificial Intelligence models.

1 Highly recommended reading is this Forbes article explaining how the development of autonomous vehicles contributed greatly to the development of synthetic data.

2. WHAT ARE SYNTHETIC DATA USEFUL FOR?

Synthetic data have multiple applications and are particularly useful in situations where the availability of real data is limited; their use requires **protecting the privacy** of the individuals involved. In the following, we illustrate **the potential applications** of synthetic data with three concrete examples or use cases:

SCIENTIFIC RESEARCH

Synthetic data allows researchers to explore and develop new approaches, models, and algorithms without the need for access to sensitive real data. This **speeds up research and fosters collaboration, while maintaining the integrity and privacy** of study participants. For example, genomic data is one of the most complex, multidimensional, and information-rich data types in the world. At just over 3 billion base pairs long, each human being's unique DNA sequence largely defines who we are, from our height to the colour of our eyes to our risk of heart disease or substance abuse. While not a natural language, **genomic sequences are textual data**; everyone's DNA sequence can be encoded through a simple four-letter "alphabet". Analysis of the human genome with cutting-edge AI allows researchers to develop a deeper understanding of disease, health and how life itself works. But this research has been limited by the **very limited availability of genomic data**. Strict privacy regulations and data sharing restrictions around human genetic data impede researchers' ability to work with genomic datasets at scale. Synthetic data offers a potentially revolutionary solution: it can **replicate the characteristics and patterns of real genomic datasets** while circumventing data privacy concerns, as it is artificially generated and does not correspond to any individual in the real world.

SOFTWARE TESTING AND SYSTEM TESTING

Synthetic data **is widely used to test and validate software and computer systems is widely used to test and validate software and computer systems**. By generating realistic but synthetic data sets, developers can simulate various scenarios and assess the performance, scalability, and security of their applications without exposing real data or taking unnecessary risks. For example, **in the development and testing of a machine vision quality control system** (in a capital goods production line), it is easier to artificially generate 100,000 images from, say, smartphones, than to have to collect those images in the real world one by one. To get real data from this application, we need processing time, sophisticated computer vision systems and a not inconsiderable number of man-hours to set up the system. With the synthetic data **we can artificially generate images** that allow us to reproduce characteristics or defects in the goods produced.

TRAINING OF ARTIFICIAL INTELLIGENCE MODELS

synthetic data is essential in training and improving machine learning models. For example, **collecting real-world driving data for every scenario an autonomous vehicle might encounter on the road would simply be impossible**. Given the unpredictability and limitlessness of the world, it takes literally hundreds of years of real-world driving to collect all the data needed to build a truly safe autonomous vehicle. In this context, companies dedicated to the development of autonomous driving systems created sophisticated simulation engines to synthetically generate the required volume of data and efficiently expose their AI systems to different driving scenarios. An example to illustrate this use can be the business activity carried out by Waabi, a company dedicated to creating synthetic simulation data on autonomous driving scenarios.

In this section we have highlighted only three concrete and representative examples to illustrate the importance of synthetic data. However, there are many application cases where synthetic data are essential, such as fraud detection, preventive medicine, credit risk assessment, training of language models, claims management in the field of insurance, etc.

3. WAYS TO GENERATE SYNTHETIC DATA

Synthetic data are generated using techniques and algorithms designed to **mimic the characteristics and distributions of real data**, without containing personal or sensitive information. Below are some common approaches used to generate synthetic data.

How is synthetic data generated?

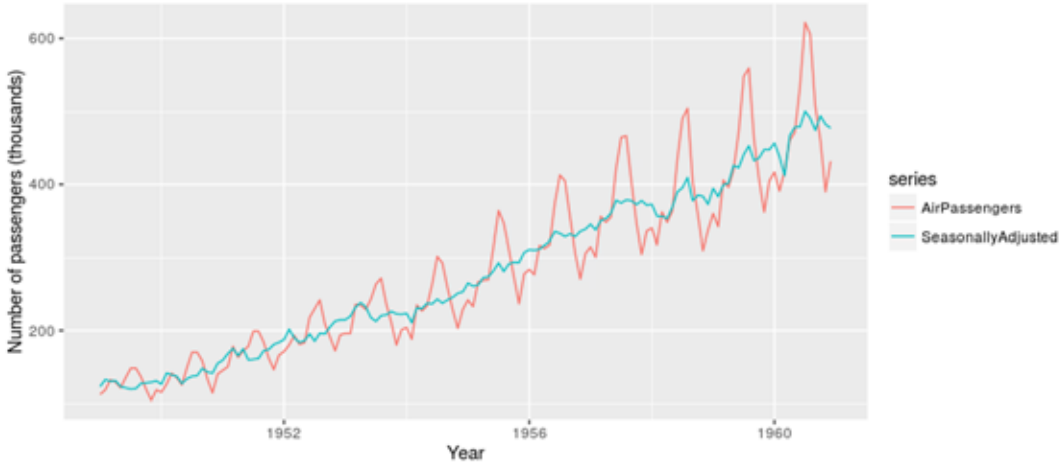
Resampling techniques: Rearranging existing datasets to obtain new ones.

Probabilistic and generative modelling: Creating datasets from probabilities obtained from real datasets.

Perturbation and masking methods: Modifying real datasets to protect privacy while preserving some features.

RESAMPLING TECHNIQUES

Using this technique we try to extract part of the original data and reorganise it randomly to **create new datasets**. This partial, random selection of numbers from a distribution is a common method of creating synthetic data. However, while this method does not capture all the information in the real-world data, it can produce a distribution of data that closely resembles the real data. A sample of this technique, although with a different purpose, is the deseasonalisation of data. For example, to obtain unemployment data that are highly correlated with particular times of the year, resampling is used to extract the underlying trend (green line) from the seasonal data (red line).



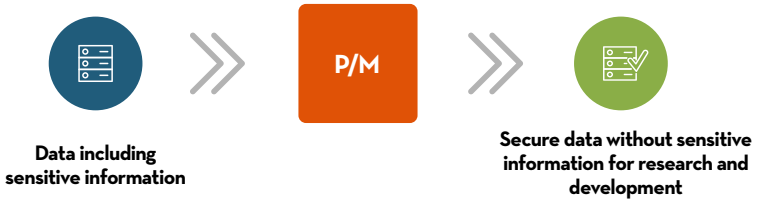
Source: StackOverflow

PROBABILISTIC AND GENERATIVE MODELLING

In this approach, probabilistic models are built based on the distributions and relationships observed in real data. These models can be statistical or machine learning models. Models are trained using real data, and then used to generate new synthetic data. Examples of generative models include generative antagonistic networks (GANs) and so-called autoencoders.

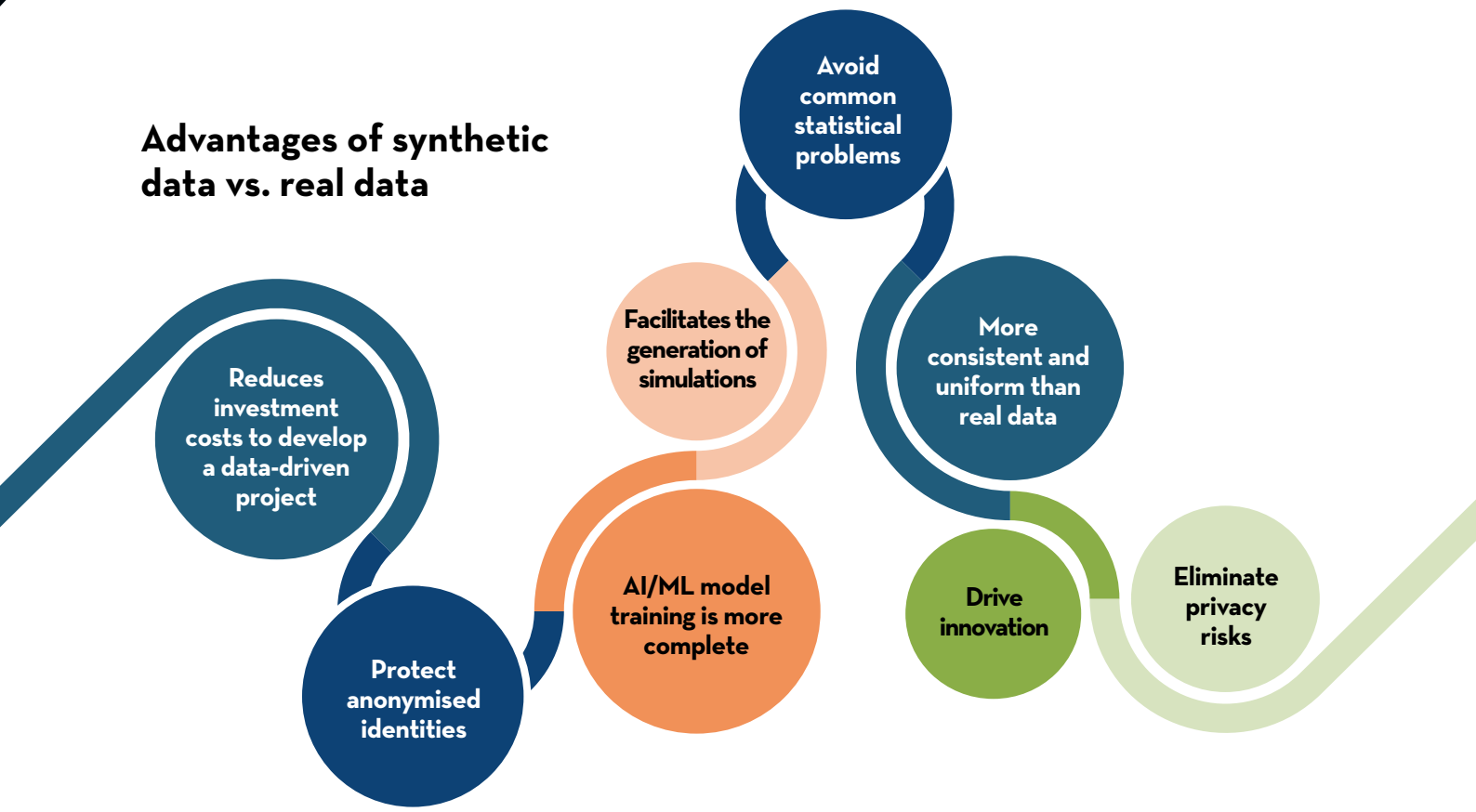
DISRUPTION AND MASKING METHODS

In this approach, the actual data are modified and disrupted in a controlled manner to protect privacy while preserving certain important characteristics. As shown in the diagram, this may involve masking sensitive information, such as replacing real names with fictitious names or distorting numerical values to avoid direct identification.



4. Benefits of using synthetic data

Advantages of synthetic data vs. real data



Throughout the report, we have already noted some of the main motivations for using synthetic data. Let us now delve into the main benefits of synthetic versus real data:

1 **Overcoming regulatory constraints.** Synthetic data avoids the regulatory constraints of real data. They can replicate all the important characteristics of real data without exposing the real information, which eliminates obligations on privacy regulations.

2 **Preservation of privacy.** Synthetic data solves the dilemma between privacy and utility because there is no need to protect synthetic data against attacks or leaks because it is not real data. Therefore, using synthetic data is a decision to use a useful dataset without putting vulnerable information at risk and maintaining data privacy.

3 **Resistance to re-identification.** A 2016 study showed that, after just 15 minutes of recording a driver's braking patterns, researchers were able to identify that driver with 87% accuracy. It turns out that the way the brake pedal is pressed is almost completely unique to an individual. Techniques exist to re-identify individuals even with anonymised real data. However, purely synthetic data do not contain real information and therefore cannot be identified.

4 **Facilitating innovation and monetisation.** Synthetic data are generally quick and easy to generate (not in all cases, but for example in the case of tabular data). As synthetic data does not present privacy concerns, it is possible to quickly share these datasets with third parties for research and innovation, and even use them as a monetisation tool.

5 **Streamlining simulation.** Synthetic data allows the generation of data that simulate conditions that have not yet occurred in real life. When real data is not available, synthetic data is the only solution; for example, the case discussed above of datasets representing possible scenarios in autonomous driving situations. This video explains such a situation perfectly.

6 **Avoid statistical problems.** Synthetic data are immune to some common statistical problems, such as non-response to elements, skip patterns and other logical constraints. By carefully designing the rules for generating synthetic data, common statistical problems can be avoided.

7 **Achieve greater consistency.** Synthetic data tend to be more uniform and consistent than real data, making them more suitable for accurate analysis. Conversely, it is true that some synthetic data are of low fidelity and may not contain the outliers or gaps that characterise real-world data.

8 **Facilitate model training and enable easy manipulation.** Synthetic data can enrich and complement real data to help train AI/ML models, especially when there is insufficient real data due to privacy, regulation and/or lack of access or time needed to capture real-world events.

9 **Making start-up projects viable and increasing profitability.** When we start a new data-driven project, we may not have had the time to capture the data or we may not even have the financial resources to buy real, quality data sets. With synthetic data we could invest a small amount in extracting a real pattern and then generate a much larger amount of data using data synthesisers. Lower investment costs increase profitability, as well as providing a viable strategy to grow the project, product or service we are building.

A PRACTICAL EXAMPLE

In this report, we have discussed various methods of generating synthetic data, such as resampling, generative models and masking of real data. To illustrate the generation of real data we will use an open source software available here that comes from a development carried out in the academic environment of MIT (Massachusetts Institute of Technology). The project is called SDV (Synthetic Data Vault) and is a Python library designed to be a complete tool for creating synthetic tabular data. The VDS uses a variety of machine learning algorithms to learn patterns from your real data and emulate them in synthetic data. It therefore belongs to the class of **generative methods for synthetic data**. The commercial version of the product is distributed through a newly created company called DataCebo. A very interesting detail of this project is that it has several practical tutorials that can be run on Google Colab.

Below, we break down the main keys with a practical example of generating synthetic data on (fictitious) customers staying in a (fictitious) hotel using Colab. All the materials used in this practical example, including the Jupyter notebook for this example to run on Google Colab, are available in the Github code repository at datos.gob.es

If you wish to replicate the exercise, we leave access to:

- GitHub repository: https://github.com/datosgobes/Synthetic-data/tree/main/Google_Colab
- Google Colab section: https://colab.research.google.com/drive/1Uo2PbmVPO4_evlbCvqwUM1c7gy-360sOv?usp=sharing

Let us begin.

The workflow we will follow during the example is like the one illustrated in the figure below:

1 **As a starting point,we use a real dataset from which we will generate our new synthetic data** that will maintain the properties and distributions of the original dataset without containing real private data. To generate synthetic data, in addition to a set of real data, we need so-called metadata. Metadata is nothing more than a description of the original dataset. That is, a characterisation of its columns (in this case because it is tabular data), describing in each case what type of data populates that column (or field). For example, in the case that the original data contains the age of the hotel guests, this will be a numeric, integer field with a range between 0-120.

2 Once we have the actual data and associated metadata, we **create and train the synthesiser.**

3 The synthesiser is basically **the programme that will create fictitious data from real data** using the probabilistic and generative modelling technique.

4 Since the actual data may contain sensitive and private data such as names, addresses, account numbers, etc., the **SDV package provides us with data anonymisation tools to detect those sensitive data and remove them, mask them, etc.**

5 **As a final step,** it is time to **check the quality of the generated synthetic data** and demonstrate that they do indeed maintain the properties of the real set but are now completely invented.

Detailing point by point, **in step 1** we can see the actual data types that make up our starting data set:

index	guest_email	has_rewards	room_type	amenities_fee	checkin_date
0	michaelsanders@shaw.net	false	BASIC	37.89	27 Dec 2020
1	randy49@brown.biz	false	BASIC	24.37	30 Dec 2020
2	webermelissa@neal.com	true	DELUXE	0.0	17 Sep 2020
3	gsims@terry.com	false	BASIC	NaN	28 Dec 2020
4	misty33@smith.biz	false	BASIC	16.45	05 Apr 2020

Actual data on customers staying in a hotel.

Associated with these, we have the corresponding metadata:


```
"METADATA_SPEC_VERSION": "SINGLE_TABLE_V1",
"primary_key": "guest_email",
"columns": {
        "guest_email": {
                "sdtype": "email",
                "pii": true
        },
        "has_rewards": {
                "sdtype": "boolean"
        },
        "room_type": {
                "sdtype": "categorical"
        },
        "amenities_fee": {
                "sdtype": "numerical",
                "computer_representation": "Float"
        },
        "checkin_date": {
                "sdtype": "datetime",
                "datetime_format": "%d %b %Y"
        },
        "checkout_date": {
                "sdtype": "datetime",
                "datetime_format": "%d %b %Y"
        },
        "room_rate": {
                "sdtype": "numerical",
                "computer_representation": "Float"
        },
        "billing_address": {
                "sdtype": "address",
                "pii": true
        },
        "credit_card_number": {
                "sdtype": "credit_card_number",
                "pii": true
        }
}
```


In step 2, what we do is provide the actual data and metadata to the synthesiser, the programme in charge of creating new data from an actual set. In Colab's notebook we found the code snippets that will run this part:

```
from sdv.lite import SingleTablePreset

Synthesizer = SingleTablePreset (
        metadata,
        name = 'FAST_ML'
)
```

After this 'training', the synthesiser now has real data references, and is ready to 'produce' them, i.e., we can generate as much synthetic data as we wish from the pattern (real data and metadata) provided.

For example, with the code execution (**step 3**) presented below, we generate a new synthetic set of 500 records:

```
Synthetic_data = synthesizer.sample (
        num_rows=500
)

Synthetic_data.head()
```

And the last line of code is used to check the result by displaying the first records of the synthesised dataset:

	guest_email	has_rewards	room_type	amenities_fee	checkin_date	checkout_date	room_rate
0	dsullivan@example.net	False	BASIC	9.436498	20 Mar 2020	11 Apr 2020	141.635838
1	steven59@example.org	False	BASIC	20.158516	20 Jun 2020	11 Aug 2020	185.529627
2	brandon15@example.net	False	BASIC	22.907020	16 Apr 2020	11 Apr 2020	145.403493
3	humphreyjennifer@example.net	False	BASIC	25.121149	04 Jun 2020	17 Jun 2020	180.463870
4	joshuabrown@example.net	False	BASIC	21.185741	11 Nov 2019	25 Oct 2019	180.288810

In the original dataset, we had some columns with data, such as the customer's email, billing address and phone number (marked by the metadata "pii": true). In the confidential synthetic data (**step 4**), these columns **are fully anonymised**[2]: they contain false values following the original format. This process has been possible thanks to the anonymisation tools of the SDV package which allows to detect sensitive data in order to remove or mask them. We can see the result with the following lines of code:

```
sensitive_column_names = [ 'guest_email', 'billing_address', 'credit_card_number']

real_dat[sensitive_column_names].head(3)
```

	guest_email	billing_address	credit_card_number
0	nichaelsanders@shaw.net	49380 Rivers Street\nSpencerville, AK 68265	4075084747483975747
1	randy49@brown.biz	88394 Boyle Meadows\nConleyberg, TN 22063	180072822063468
2	webermelissa@neal.com	0323 Lisa Station Apt. 208\nPort Thomas, LA 82585	3898347697 1380

Conjunto de datos original (real)

	guest_email	billing_address	credit_card_number
0	dsullivan@example.net	90469 Karla Knolls Apt. 781\nSusanberg, NC 28401	516103375951898 3
1	steven59@example.org	1080 Ashley Creek Apt. 622\nWest Amy, NM 25058	413304741314547 5690
2	brandon15@example.net	99923 Anderson Trace Suite 867\nNorth Haley, T...	4977328103788

Conjunto de datos sintético (ficticio)

Finally, in **step 5** we evaluate the quality of the synthetic data generated. It is important to **verify that the data generated maintain the properties of the original set**, since this will determine whether the systems, processes, and algorithms that we subsequently train will behave as expected.

For this, the SDV package provides us with a series of **specific quality control tools** that facilitate the work. For example, the following code generates a small quality report:

```
from sdv.evaluation.single_table import evaluate_quality

quality_report = evaluate_quality(
real_data,
synthetic_data,
metadata
)
```

With the following result:

```
Creating report: 100%|██████████| 4/4 [00:00<00:00, 27.02it/s]
Overall Quality Score: 89.12%

Properties:
Column Shapes: 90.27%
Column Pair Trends: 87.97%
```

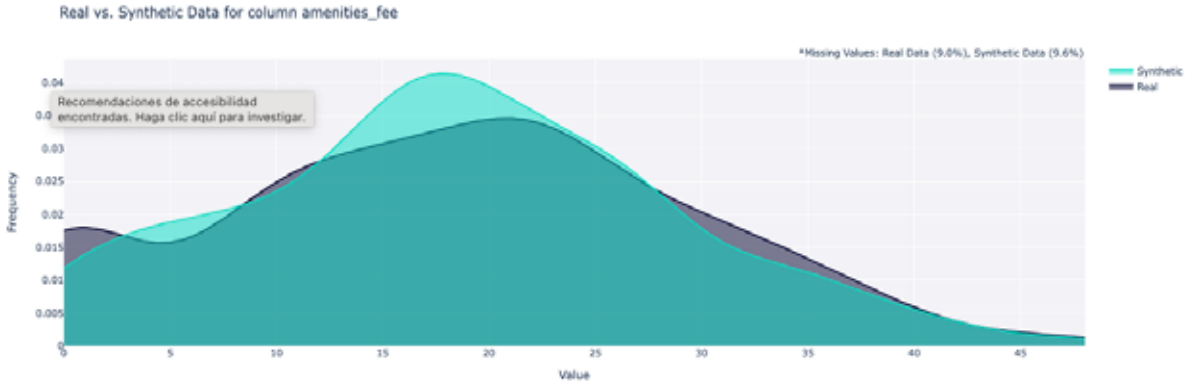
[2] See the Guide to Data Anonymisation: Techniques and case studies: https://datos.gob.es/en/documentacion/introduction-data-anonymisation-techniques-and-case-studies

The quality report assesses the ability of the generated synthetic data to capture the mathematical properties of the real data. This is also known as **synthetic data fidelity reporting**. The report runs selected metrics, such as **the mathematical distribution of values in a column or the correlation between a column of synthetic data and its corresponding column of real data.**

For example, the metric, Column Shapes, assesses the similarity between the mathematical distribution of a synthetic column versus its real one. For example, let's imagine that we are generating synthetic data corresponding to a set of students' marks in an exam. If we assume that the distribution of marks in an exam follows a normal or Gaussian bell distribution, a good synthetic column should reproduce the same distribution. This is what the Column Shapes metric measures. The higher the value of this metric, the more similar the distributions of the two columns are. The overall metric is the average of the individual metrics for each pair of columns. In the case of the Column Pair Trends metric, the mathematical correlation between a synthetic column and its corresponding original is evaluated. Recall that a maximum correlation value of 1 means that the points plotted in one column against the other draw a perfect straight line of slope equal to 1. The same applies to a negative correlation (anti-correlation) but with opposite sign. Further information on the tool's quality reports can be found in the following documentation: https://docs.sdv.dev/sdmetrics/reports/quality-report/whats-included.

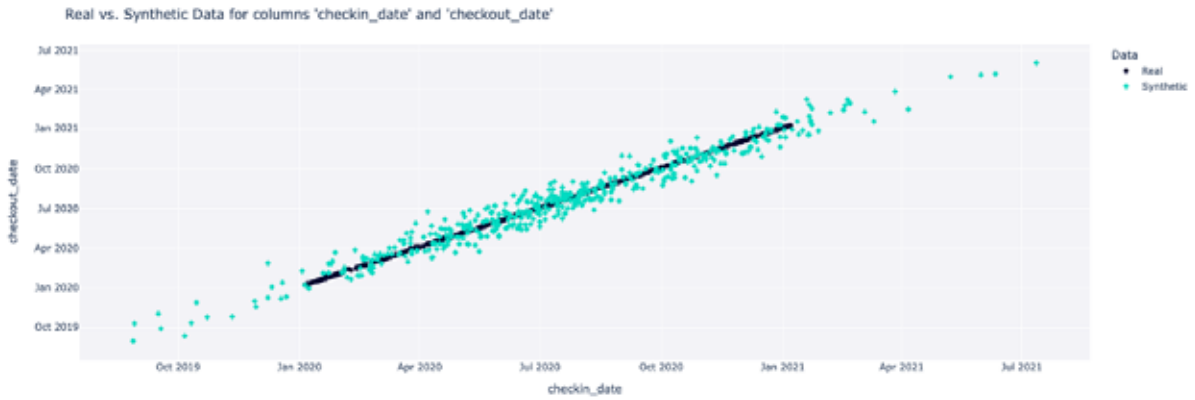
The quality report is also available in a more graphical and visual form. The following code generates a visualisation that allows us to visually represent the Column Shapes metric (real vs. synthetic) to visually check that they maintain the same statistical distribution.

```
fig = get_column_plot(
    real_data=real_data,
    synthetic_data=synthetic_
data,
    column_name='amenities_fee',
    metadata=metadata
)

fig.show()
```



Complementarily, we can see the correlation (remember that it is measured by the Column Pair Trend metric) of certain variables, such as the hotel check-in and check-out dates of both sets:

```
fig = get_column_pair_plot(
   real_data=real_data,
   synthetic_data=synthetic_data,
   column_names=['checkin_date', 'checkout_date'],
   metadata=metadata
)

fig.show()
```



- In this example, we have seen probably the simplest synthetic data generation exercise to execute. However, SDV provides us with other guided tutorials in Google Colab, which are very interesting to further deepen in a practical way in the generation of synthetic tabular data.

    - For instance, we can explore the generation of synthetic data from our own data in csv format.
    - We also have the possibility to generate synthetic data from databases with several related tables.
    - We can also try to customise the data synthesiser with some parameters and business rules, e.g. force certain data to comply with a certain range (dates in a certain period)

CONCLUSIONS

Synthetic data offer a secure and affordable alternative when real data are scarce, inaccessible or cannot be used due to privacy restrictions. This type of data is artificially generated but maintains similar properties to real data, which makes it useful in certain varieties of use cases. However, it is important to **assess the quality of synthetic data** and ensure that it meets specific requirements and objectives to avoid introducing bias or unrealism before implementation.

Synthetic data is now a very present reality, especially since the explosion in late 2022 and early 2023 of generative AI technologies, massive language models and image generators. Gartner's predictions for 2024 (made in 2021) are likely to be far exceeded (60% of data will be synthetic) and we are likely to face new challenges related to synthetic data, such as its lack of precision, its dependence on real sets, the risk of not taking into account anomalous or exceptional situations and the lack of transparency towards consumers due to not knowing whether the data is real or fictitious. It is impossible to predict what the future holds for synthetic data, especially with the **exponential development of artificial intelligence technologies**, but what is clear is that it is going to play a crucial role for all of us, consumers, professionals, academics, students, etc. See you in the next report on data, synthetic or real, who knows?

GOBIERNO DE ESPAÑA
VICEPRESIDENCIA PRIMERA DEL GOBIERNO
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL
red.es
Oficina del Dato
Iniciativa aporta
datos.gob.es