**Synthetic Data Generator FAQ**

**Contributor: Siris Neupane**

**GitHub Issue Number: #155 (picked from the backup issue because my issue (#153) was solved by someone else)**

**What is synthetic data?**

Synthetic data is like a clone of real-world data but made from scratch without using any actual real data. It imitates the patterns and characteristics of real data, which means it behaves similarly for testing or training purposes in various applications but doesn't include any real life sensitive information.

**This Synthetic Data Generator project is a powerful tool for companies and researchers needing to work with secure, non-sensitive data. It provides a robust framework for synthetic data creation that supports privacy, compliance, and innovation in an increasingly data-centric world.**

**One of the major takeaway was understanding the crucial role of metadata in synthetic data generation. By refining how we handle metadata, including supporting various data types and enhancing automatic data type inference, we greatly increased the accuracy and utility of the synthetic data produced.**

**What are the differences between synthetic data and real data?**

The main differences are:

* **Source**: Real data comes directly from the real world, capturing actual events and behaviors. Synthetic data is generated by algorithms and doesn't come from real-world events.
* **Privacy**: Synthetic data is designed to not include sensitive information, making it safer for use in environments where data privacy is a concern like healthcare, financial services, etc.
* **Flexibility**: Synthetic data can be created to match specific scenarios or conditions without the limitations or ethical concerns that come with collecting more real data.

**Why do we need to use synthetic data?**

Synthetic data is super useful because it:

* Protects privacy by not using real personal data.
* Allows developers and researchers to test systems thoroughly without waiting for new real data to be collected.
* Helps in situations where using real data can be risky or restricted by law.

**In what situations should synthetic data be used instead of real data?**

Synthetic data is particularly helpful:

* When you can't use real data because of privacy laws or ethical issues.
* In testing and development phases of software, especially when the software handles sensitive information.
* When you need data that is hard to collect in the real world, like rare events or scenarios.

**How to ensure the privacy and security of synthetic data? How to prevent synthetic data from being misused or leaked?**

To keep synthetic data secure and private:

* Use techniques like differential privacy during the data generation process to ensure the data cannot be traced back to any real individuals.
* Apply strict access controls and encryption to store and manage synthetic data safely.
* Regularly audit and review how synthetic data is used and generated to make sure it remains compliant with all privacy guidelines and regulations.

**SDG uses some smart models like LLM and CTGAN to make this fake data look really convincing. This means businesses can test their systems, train models, and share data without worrying about breaking privacy rules. It’s kind of like using a stunt double in movies; it keeps the real actors safe while making everything look real.**