6.1: Sourcing Open Data

Mohammed Molla

**Data Source:**

The SAML-D dataset is a structured financial transaction repository designed for anti-money laundering (AML) research. It comprises 9,504,852 transactions.

This dataset addresses key AML challenges, including class imbalance, behavioral pattern masking, and cross-border transaction complexity, making it valuable for both research and industry testing.

This dataset is sourced from Kaggle and can be accessed from here ([Anti Money Laundering Transaction Data](https://www.kaggle.com/datasets/berkanoztas/synthetic-transaction-monitoring-dataset-aml/data)) for reference.

**Data Collection:**

The dataset, named SAML-D, was synthetically generated using a custom-built AML transaction generator. This generator was developed based on insights from existing datasets, relevant literature, and semi-structured interviews with AML specialists.

The aim was to create a comprehensive dataset that includes a diverse range of transaction typologies, geographic locations, high-risk countries, and high-risk payment types, which are often underrepresented in real-world data due to privacy and legal constraints.

**Data limitations:**

Due to it being a synthetic dataset, it will always come with potential limitations.

1. Even though it's designed to mimic real financial transactions, synthetic data can’t fully capture the unpredictability and nuance of real-world behavior—especially with adaptive, evolving criminal activity in money laundering.
2. Since part of the dataset is generated using expert-designed rules and typologies, it might reflect the biases or limitations of current AML knowledge. This could lead to:
   1. Overrepresentation of known fraud types.
   2. Underrepresentation of novel or subtle laundering patterns.
3. Synthetic data lacks the true variability in human behavior. While patterns may be simulated, real customer anomalies or context-driven decisions (e.g., sudden transfers due to emergencies) are difficult to replicate.

**Reason for this Data selection:**

In the USA alone an estimated $300 billion is laundered annually. Digitalization and the rise of cryptocurrencies are contributing factors, as criminals exploit new technologies to launder money more effectively and evade detection.

This dataset aligns closely with my research interest in applying data analysis techniques to address money laundering challenges within the financial sector. Through this project, I aim to contribute meaningful insights by leveraging my analytical skills, while also further developing my expertise as a data analyst in a real-world, impact-driven context.

**Data cleaning summary:**

**Consistency checks:**

* Checked for missing values in all columns
* Checked for duplicate values in all columns
* Checked for mixed types for all columns
* Checked for descriptive statistics on all numerical columns

**Data wrangling:**

Dropped columns ‘Is\_laundering’ and ‘Laundering\_type’

**Save Cleaned dataset:**

Exported cleaned dataset to prepared data file directory for analysis

**Data ethics:**

**Privacy & Data Anonymization:**

‘Sender\_account’ and ‘Receiver\_account’ colums contain account numbers, which could be personally identifiable information (PII).

* Ethical Risk: If these are real account numbers, they could be used to trace individuals, violating privacy laws.

**Bias:**

Geographical Bias: Over-representation of certain countries (e.g., UK transactions dominate).

Currency Bias: Some currencies (e.g., UK pounds) may be flagged more often than others.

**Define questions to explore:**

**1. Temporal Trends & Transactional Behavior**

* What are the peak times (Time and Date) for high-value transactions?
* Which sender-receiver pairs have the most frequent transactions (potential smurfing or structuring)?

**2. Risk Assessment & Compliance**

* Which banks/locations receive the highest volume of suspicious transactions?
* Are certain payment methods (e.g., Cash Deposits vs. Wire Transfers) riskier for fraud?

**3. Financial & Operational Efficiency**

* What is the average transaction amount per payment method?
* Which currency pairs are most common in cross-border transactions?

**4. Geographic & Cross-Border Analysis**

* Which countries are most involved in high-risk transactions?
* Are there corridors with unusually high transaction volumes?