# Data Overview

While this is sample data and not directly extracted from a client database, it was simulated to follow and reflect patterns in client data, so we ask you do not distribute these CSVs.

## AML\_sample\_transactions.csv

This is where a great deal of your “predictive” data is stored. This csv tracks the transactions that occurred within the bank over a given time period. Below are overviews of the most relevant columns.

* Txn\_id: unique identifier for the transaction
* Txn\_timestamp: time of the transaction
* Txn\_type: Wire, ATM, and ACH are the most common types of transactions recorded in this data
* Txn\_amount: the amount given/received in the transaction
* Customer\_id: the customer conducting the transaction
* Customer\_acct\_id: the account conducting the transaction
* Debit\_credit: Flag to indicate if you are receiving incoming funds (credit) or giving outgoing funds (debit)

Note there is an id for the customer and the account. A customer can be on multiple accounts, and an account can contain multiple customers. There is merit in aggregating to either level or evaluating the combination of the two.

## AML\_sample\_accounts.csv

If you chose to aggregate features to the account level, this is the descriptive information about each account. While you may not find it useful, we wanted to provide this dataset for completeness.

## AML\_sample\_individuals.csv

If you chose to aggregate features to the customer level, this is the descriptive information about each customer. While you may not find it useful, we wanted to provide this dataset for completeness. Please ignore the nameAlias column – it is used for a different purpose. Obviously, proceed with caution regarding including demographic data in a model, especially in a highly regulated space.

## AML\_sample\_accountMapping.csv

This shows which customers are tied to which accounts.

## AML\_sample\_alert\_scored.csv

This is where you will derive your response variable (and potentially additional explanatory variables depending on the perspective of the team).

* Alert\_id: unique identifier for an alert that was raised by the primary monitoring system that monitors behavior and raises alerts when a rule is violated
* Cust\_id/acct\_id: the customer/account relevant to the alert
* Start\_ts: the time of the alert
* Priority: this is an urgency level that is determined and assigned by the primary monitoring system
* Alert\_score: this is a score of how likely the alert is to be escalated, determined by the primary monitoring system. It may be interesting to try including and not including for sensitivity analysis, but do not fully trust this score.
* Disposition: this is the ground truth of the alert’s eventual outcome. The three categories are: SAR, Escalated, Waived.
* FP\_TP\_Tag: this is a column that is sometimes used as the response variable. It is a flag of FP=False Positive and TP=True Positive. Here, FP means that the primary monitoring system incorrectly created an alert, and TP means that the primary monitoring system correctly created an alert. This is a column our team created by saying: if the disposition was SAR or Escalate then TP, otherwise Waived. You do not have to use this column or follow this logic.

## AML\_sample\_alerts\_5.csv

These are 5 alerts that we hold separate from the others. We use them as “new and incoming” alerts to feed into our entire pipeline (the predictive models being one part). For the most part they have the same columns as the alert\_scored csv. Feel free to use them as a sanity and operations check, but valuable data will be found in the alert\_scored csv.