Dataset Documentation

The rows represent each individual person and the columns are their respective attributes. There are duplicates of people in the dataset who have varied amounts of noise introduced to certain fields. The goal is to generate hashes and properly be able to identify the same person across all their duplicates in this dataset, among others.

**Rows**: 2,064,313

**Columns**: 12

Columns

**ID**: Indexer used for processing dataset, not of relevance or use for hashing

**SEX**: String, Gender of the person, value is either ‘M’ for male or ‘F’ for female

**FIRSTNAME**: String, first name of person, no null values

**MIDDLENAME**: String, middle name of person, possibility for null values (i.e. no middle name)

**LASTNAME**: String, last name of person, no null values

**COB**: String, name of city of where the person is from

**DOB**: Integer ranging from 1 to 31, day of birth for given person

**MOB**: Integer ranging from 1 to 12, month of birth for given person

**YOB**: Integer ranging from 1948 to 2000, year of birth for given person

**CNOB**: String, country of birth for given person

**SUBJECTHASMIDDLENAME:** Boolean, True if person has a middle name, if not, False

**SYNTHETIC\_SSN**: String, 9 digit GIID value resembling the SSN. It is unique for each given person and duplicate people have the same SSN.