## Overview of Data

We utilized data we found out on kaggle.com related to loan defaults. There were a variety of options available to us (figure?) but we decided to focus on two main sources of data. Application\_train, which contained a set of 122 columns representing various demographic and other factors related to an individual’s specific credit application, and bureau, which corresponded to credit bureau related debts associated with the user represented by the aforementioned credit application and had 17 columns. These tables are related in a 1:n (application\_train to bureau) fashion with a common key of SK\_ID\_CURR.

The first task we did with respect transformation of the data was to consolidate and flatten this relationship. This was primarily done to make our models less complex structurally without needing to sacrifice relevant data. We accomplished this via storing the respective csv files in a SQLite DB and then joining them on SK\_ID\_CURR. We decided to use only two fields from the bureau file/table and those were AMT\_CREDIT\_SUM and AMT\_CREDIT\_SUM\_DEBT. Because of the 1:n, we summed the fields respectively to produce one number for both columns for the respective SK\_ID\_CURR which was our unique key. Our outcome variable was a binary field called TARGET where 1 indicated a loan that had gone into default at some point and 0 a loan that always been in good standing.

### Data Transformation

Having done that, we still had many data cleaning activities to complete before a usable set of final data could be produced. We will detail these below:

1. Determine how many records within our dataset contained NA or missing values.
   1. For any kind of missing record, we found that out of 217150 initial records, 209935 had at least 1 column with NA or missing data. We knew we needed to focus on eliminating these missing values where possible.
   2. We also checked for missing data in 3 specific fields called EXT\_SOURCE\_1, EXT\_SOURCE\_2, and EXT\_SOURCE\_3. These fields corresponded with normalized values (0-1) corresponding the credit bureau scores. We found 64 records missing all 3 of these values. We would use this later
2. Next, we transformed our missing or NA data based on 5 distinct approaches
   1. For a set of columns, we determined that it was reasonable to set these values to 0.00 when missing. This was done because these fields contained information where missing data essentially meant 0. An example field would be OWN\_CAR\_AGE. If missing, we interpreted that to mean that they did not own a car and 0 was appropriate
   2. Next, there were two fields where we felt that NA and missing should be replaced by “Unknown”. OCCUPATION\_TYPE was one such.
   3. Next, there was one column, CNT\_FAM\_MEMBERS that represented the size of the family of the applicant. Since the applicant was at least 1, we set that value to 1 where missing or NA.
   4. Fourthly, and related to step 1b above, we created a new column called CREDIT\_MISSING (binary) for those rows where EXT\_SOURCE\_1, EXT\_SOURCE\_2, and EXT\_SOURCE\_3 were all missing. Those rows received a 1 where all else with at least 1 value in the 3 were set to 0. We intended to use this to try and help branch our modeling for those with or without credit
   5. Lastly for this phase, for those records missing 1 or 2 values (but not all 3), we imputed the mean respectively and added those to the missing values.
3. Next, we dealt with outliers.
   1. We capped the CNT\_CHILDREN column at 4 when a larger value was found. It was clear that there were errors in this data for some records.
   2. We also deleted a row where the reported income of the individual was over $10 million. We believed this to potentially be a mistake given the rest of that applicants data

### Dimensionality Reduction

Our next step was to attempt to reduce dimensionality due to the large set of potential explanatory variable. We ran PCA analysis and k-Means Clustering on the transformed data. This, unfortunately, was not particularly useful for reduction as the scree plot suggested 1 factor had the primary impact and that comprised most of our fields. I’ve included a graph (see figure).

Because the results of PCA and k-Means did not effectively reduce our dimensionality, we decided to use our best guesses and expertise to remove fields that were either not likely to contribute significantly to the outcome OR were missing enough data that the results would not possess a lot of power. An example of this was the removal of all columns that dealt with the sq foot size of the reported applicant’s residence. In the end, we removed all but 49 fields from our “final” dataset. 48 predictor fields and 1 outcome, which was describe above as TARGET.

Each team member then utilized the same dataset on the models they were responsible for evaluating.

A diagram of a company

Description automatically generated

A diagram of a clustering graph

Description automatically generated with medium confidence