CSE 801 Project

Data Cleaning Notes

* Original data contains 2260701 observations and 151 variables.
  + High dimensionality and number of observations.
* Data contains loans issued in the period of 2007-2018.
  + We are going to work with only the last three years (loans issued in 2016-2018.
* Some of the variables are better described in the dictionary of the information. We did a match with this table to get a better description of the variables. In this way we can eliminate the variables that are not relevant.
  + We pick only the variables that were in both tables (data and dictionary)
* We fixed the format of the variables containing dates.
* We transformed the ‘emp\_length’ variable for it to be numeric
* Handling NA values
  + For 'emp\_title', 'verification\_status\_joint', we filled them with ‘’.
  + For 'bc\_open\_to\_buy', 'mo\_sin\_old\_il\_acct', 'mths\_since\_last\_delinq',
  + 'mths\_since\_last\_major\_derog', 'mths\_since\_last\_record',
  + 'mths\_since\_rcnt\_il', 'mths\_since\_recent\_bc', 'mths\_since\_recent\_bc\_dlq',
  + 'mths\_since\_recent\_inq', 'mths\_since\_recent\_revol\_delinq',
  + 'pct\_tl\_nvr\_dlq','sec\_app\_mths\_since\_last\_major\_derog' we filled them with the max value of each column
  + For the rest of the columns, we filled them with the minimum value of each column.
* The final dataset contains 938821 observations and 102 variables.

Next steps

* Data Exploration
* Multicollinearity
  + We can delete some of the columns that are too strongly correlated.