**Capstone Project-1: Data Wrangling**

**Jun Yin**

1. **Data Acquisition**

The dataset was acquired from kaggle (<https://www.kaggle.com/c/home-credit-default-risk/data>.)

There are 10 datasets from the original data source: the first two application\_train and application\_test are the two main datasets with the information about each application, sorted by the feature SK\_ID\_CURR and identified by the feature TARGET representing repaid by 0 and not repaid by 1. The bureau dataset displays the applicants’ previous credits from other financial institutions. The bureau\_balance exhibits monthly balances of previous credits in the Credit Bureau. POS\_CASH\_balance shows monthly balance snapshots of the previous point of sales and cash loans that the applicant had with Home Credit. The previous\_application tells all previous application records for Home Credit loans. And all repayment history for previous loans are recorded in installments\_payment dataset. The rest are one file for explaining each column and one submission example file

1. **Missing Value Treatment**

In this project, the first step is to check with every column name and understand the meaning of each column, and identify numeric or categorical data in each column. One of the most common problems faced in Data Cleaning/Exploratory Analysis is handling the missing values. There are several causes of missing values: sometimes values are missing because they do not exist, or because of improper collection of data or poor data entry. In that case, various filling strategies are required to operate for different situations. Here use three different filling strategies in this project:

1. Filling in with mean;
2. Filling in with most frequent value/item;
3. Filling in with 0;
4. **Feature Engineering**

In order to generate more features, supplementary datasets are merged into the main dataset. Since most of machine learning models cannot handle categorical variables, encoding is necessary to perform in this section. Here, One-hot coding is applied to transform all categorical variables into numerical variables, and all original categorical columns are dropped.

1. **Next Step**

After done with data wrangling and data preparation above, the exploratory data analysis (EDA) will be the follow step.