Home Loan Credit – Wrangling

Introduction

This is the first step to analyse the capstone project data to explore data structure, variables, data integrity, missingness and correlation between variables. There are eight datasets in total for this project – two are the main home-loan applications with target variable (train) and without target variable (test), and we analyse these two data in this stage. Other six datasets are historical records and would be interpreted with aggregation for each applicant later.

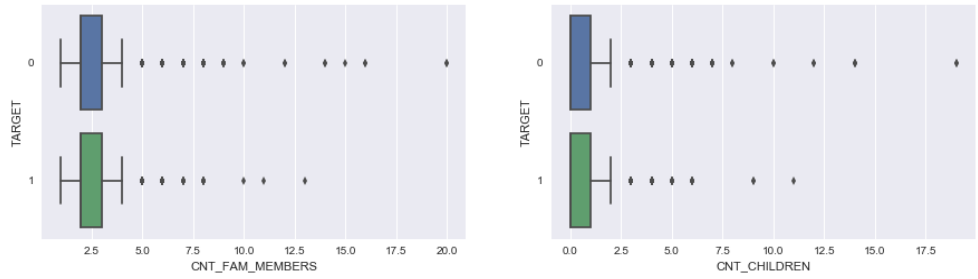
Data structure

Train data consist of 307511 rows and 122 columns and test data 48744 rows and 121 columns excluding the target variable. Because of data size, the first few rows and the last few rows have been checked – the column names are all written with capital letters consistently. However, the summary table shows that there are a large number of missing in data and abnormal values in some columns. We look at the detail of these issues in later sections. Aside, the target variable indicates a flag of the payment difficulties for each applicant (1 means 'Yes' and 0 'No'); there are 282686 zeros and 24825 ones, implying about 8.1% of applicants are flagged as 'having difficulties to pay-back'.

Outliers

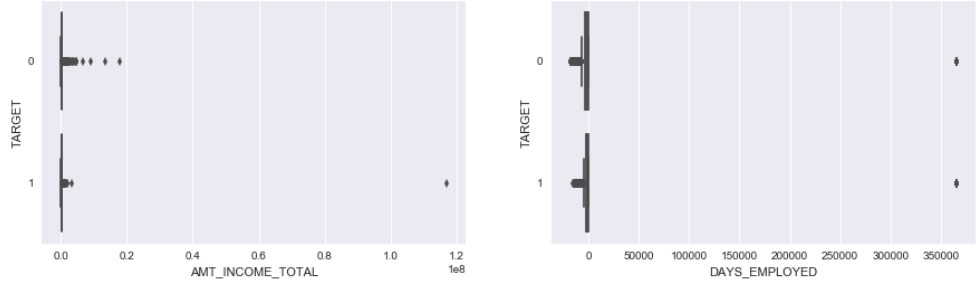
Distributions of the number of families and the number of children have very long right-tails, reaching up to 20. Boxplots in Figure 1 shows that 75% of data have the family size of less than four and less than two children, but upper 25% of data are spread from five to 20 family members and three to 19 children. It seems that there are some large-size families with many children. The maximum of 20 and 19 for the family size and the number of children, respectively, seem to be unusual, but still can happen – so we reserve those values without any implementation.

Figure 1: Boxplots of family size and the number of children by target



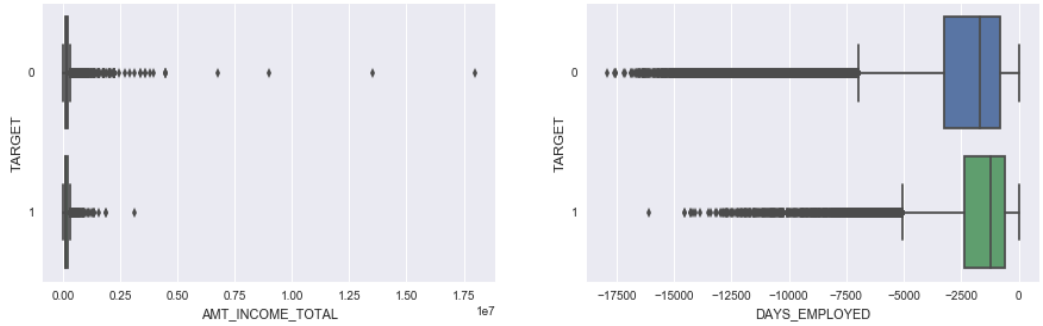
Monthly total income and days of employment variables seem to be abnormal as shown in Figure 2.

Figure 2: Boxplots of monthly total income and days of employment



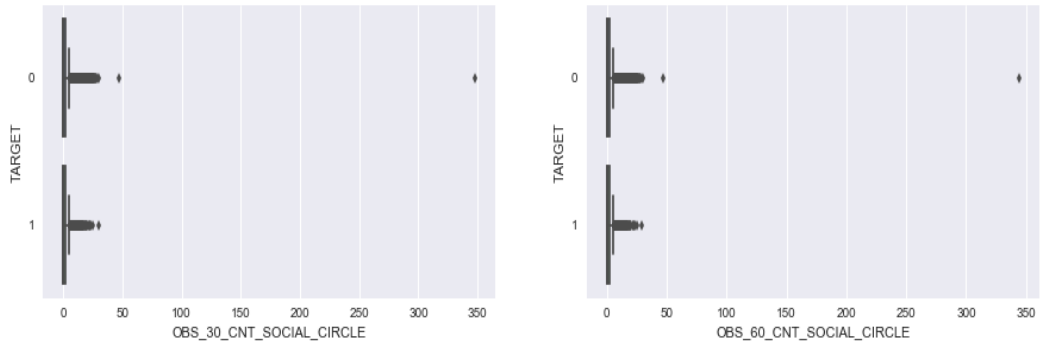
The plots show that a few outliers on the right-hand side distort the plots, and we con not see the clear features of their boxes. The maximum income of 1.17e+08 could be an error that was typed with extra zero(s); so we replace this with null. The days of employment are denoted with negative numbers representing the backward counting from the day of application - so the positive value of 365243 days defining 1000 years in future is suspicious. We found that their primary income source is the pension, so we replace those values with null. In the same manner, all columns representing days have checked and implemented for the two datasets. Figure 3 depicts the updated boxplots for these two variables, even though there are still a large number of abnormal data points.

Figure 3: Boxplots of monthly total income and days of employment after implementation



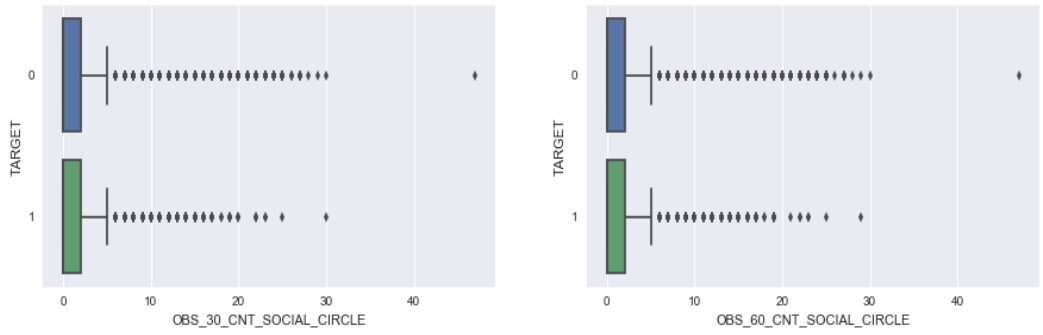
Social-circle counts for 30 days and 60 days have the similar characteristic as one in the income total as shown in Figure 4.

Figure 4: Boxplots of social-circle counts for 30 days and 60 days



One value counts 344 social circles in 60 days as well as in 30 days - it is extreme comparing the most values under 50, and we replace this with null in both columns. After the implementation, we see clearly a number of outliers as well as the boxes and whiskers on Figure 5.

Figure 5: Boxplots of social-circle counts for 30 days and 60 days after implementation



Missingness

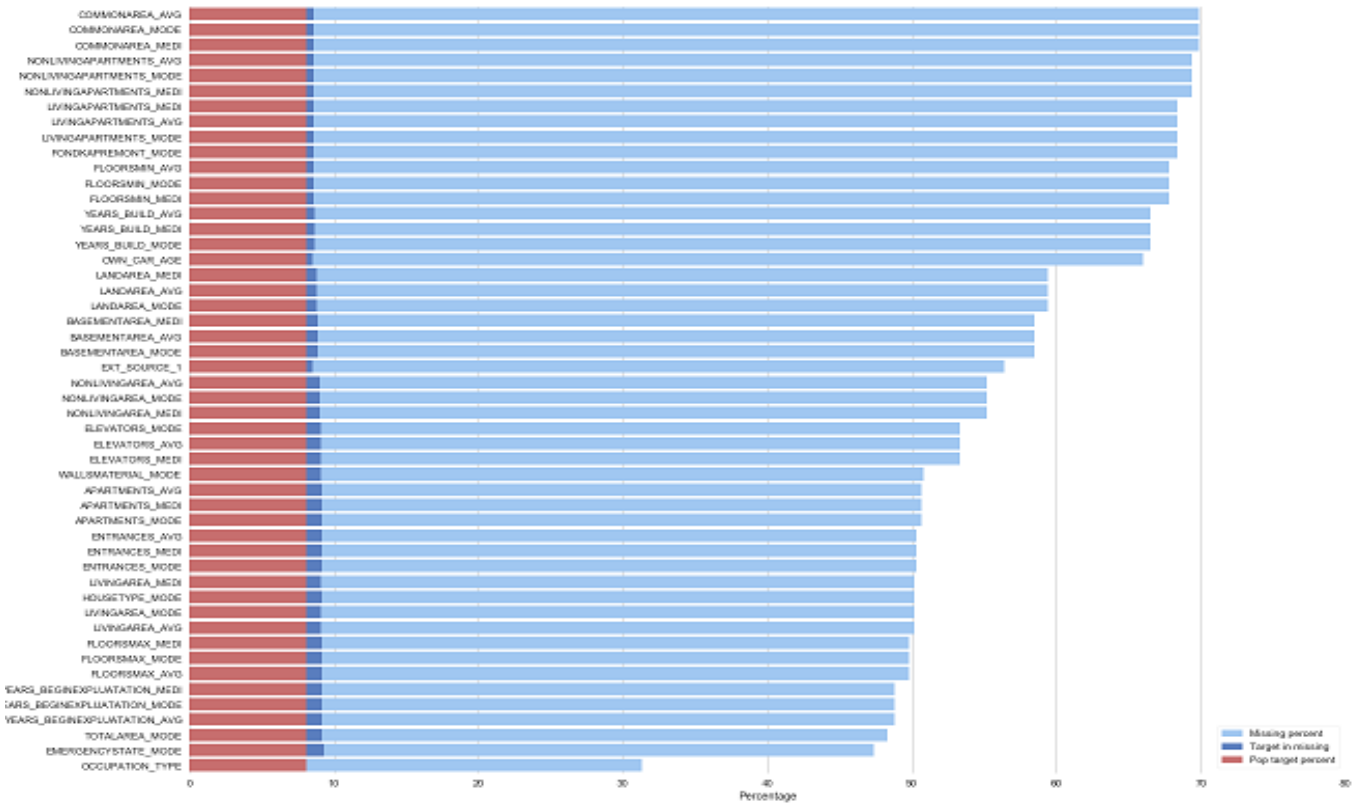
Among 121 total variables 64 have at least one missing values – the highest volume of missing is 69.9% in the normalised values of the building where the clients live and followed by other similar information with 69.4% as shown in Table 1.

Table 1: The most missing variables with the target = 1 percentages for the missing and non-missing groups

	variable	na_percent	na_target_percent	no_na_target_percent
61	COMMONAREA_MODE	69.9	8.6	6.9
75	COMMONAREA_MEDI	69.9	8.6	6.9
47	COMMONAREA_AVG	69.9	8.6	6.9
83	NONLIVINGAPARTMENTS_MEDI	69.4	8.6	6.9
55	NONLIVINGAPARTMENTS_AVG	69.4	8.6	6.9
69	NONLIVINGAPARTMENTS_MODE	69.4	8.6	6.9
53	LIVINGAPARTMENTS_AVG	68.4	8.6	6.9
67	LIVINGAPARTMENTS_MODE	68.4	8.6	6.9
81	LIVINGAPARTMENTS_MEDI	68.4	8.6	6.9
85	FONDKAPREMONT_MODE	68.4	8.6	6.9

Figure 6 shows that the 49 variables have almost 50% or more of missing, displayed by the light-blue colour bars. The dark-blue bars indicate the percentage of labelling in target = 1 for the missing group and the same-size red bars the percentage of target = 1 for the whole train dataset. This plot implies that the variables with at least 50% missing seem to have higher percentages of target = 1 labelling in missing group comparing it in the whole data.

Figure 6: Top 50 missing variables

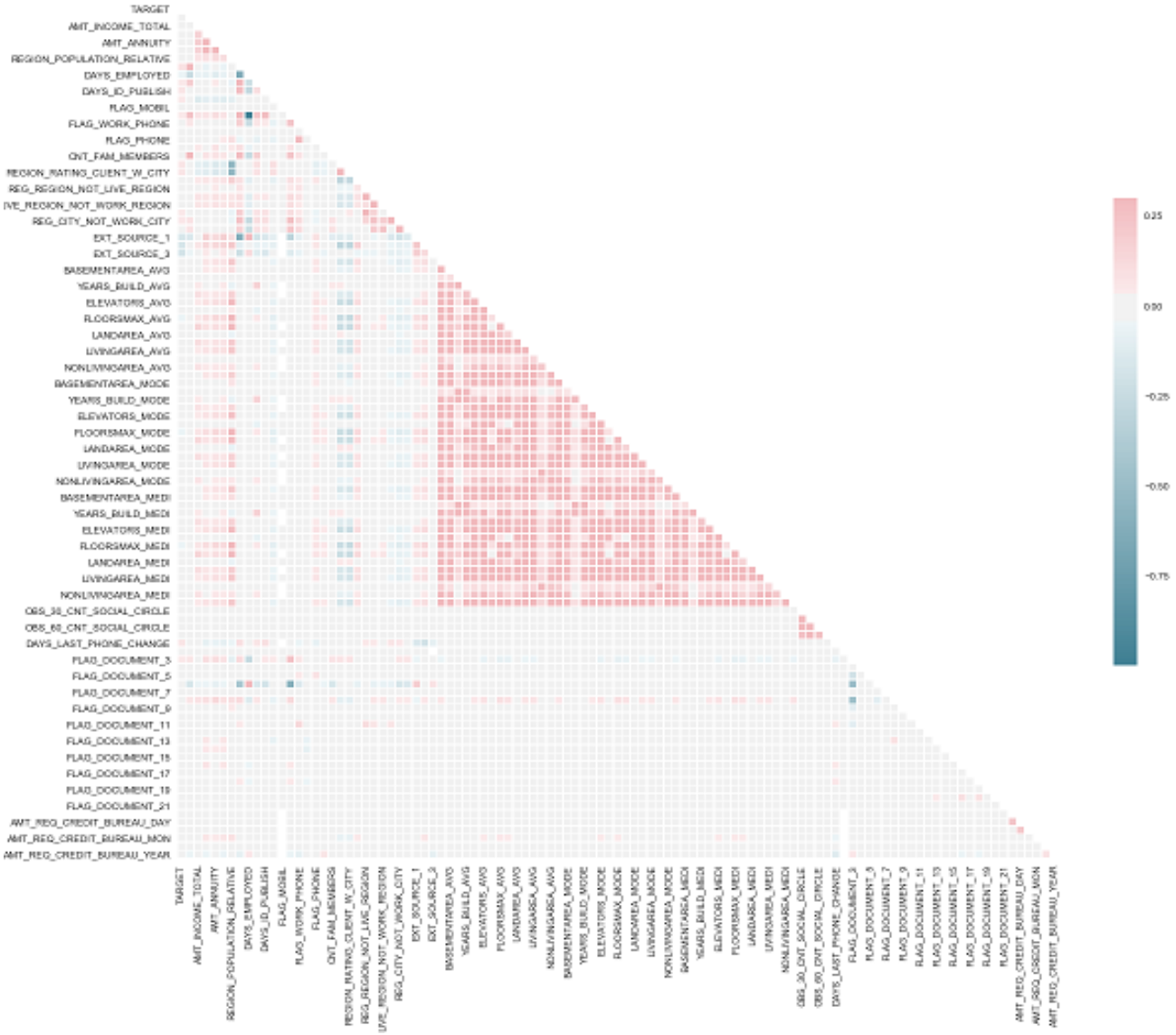


To verify the differences of target = 1 percentages between missing and non-missing groups, we computed the Chi-square test for each variable; the test result tells us that the differences between two groups are statistically significant in 95% confidence level for 62 variables among 64 having at least one missing value. We implement 62 missing indicator variables for this significant differences – each new variable has binary values, 0 for non-missing and 1 for missing.

Correlation

Correlations for each pair of the all numerical variables of train data have been computed - as the dimension of data is quite big, the correlation matrix is not representable. Instead, the heat-map shows a better insight expressing the negative or positive relationship with contrast colours and strengths with hue as shown in Figure 7. Due our major concern is the correlation with the target variable, we focus on the very first column of the heat-map: we can see just a few light hues of red (positive) and blue (negative) colours, but mostly close to white colour implying most variables have not much strong association with target variable.

Figure 7: Heat-map of the correlation matrix for the train data



We are considering to remove variables with correlation coefficients with the target of (-0.02, 0.02) for avoiding unnecessary computation time to model with a large number of predictors.