ANLY482 Meeting Minutes with Sponsor

(01 March 2018)

Date:	1 March 2018
Time:	08:00-09:00
Venue:	SMU SIS Meeting Room 5.1
Attendees:	Team: Ruiyan, Qian, Nicholas, Sponsor: Prof Kam
Agenda:	Project Overview for Kiva Dataset

S/N	**Things Discussed/Done**	**Remark**
1.	Discussion of findings	• Prepare a data dictionary to get a clearer description and definition of data. Put metadata in clearly, descriptions of the fields and the data type (binary, numerical, categorical). For categorical data include the different classes, as well as number of records and missing values. • When predicting or estimating welfare level, what are the factors affecting welfare of the people? Suppose there are 168 countries, do cross validation, 1 country as reserve and run against the other countries. Model then tells us how well can we predict the missing

		country. Alternatively, use sampling and use only 100 for sampling, and remaining 68 for validation. MPI is too aggregated, use it as a respondent variable to calibrate the model. • Clean the data first, explore by countries and regions. Look at borrowing rate across countries, some much higher than others. Might only focus on Philippines, borrowers are looked at using domestic currencies. MPI is measured against US dollar, to some extent we can go down to administrative/3rd level (villages), which can be found using latitude and longitude. Explore the data and get an idea of what to investigate. Borrower from Philippines may come from particular islands and show signs of concentration. Analyze how borrowing patterns change over time and over space. • Identify and study the distribution of the loan amount. Breakdown by country, total loan amount, number of loans, create bar chart of different countries with number of loans and sum of loans. Look at average loan amount as big population might distort data. Look into country by theme/sector, for example how many people take up loans for education, and how does it vary across continents/countries? Use ANOVA to test if differences in box plots are statistically significant. Ensure methodology is in line with objective. • Temporary objectives are loan amount, repayment period and rate of funding. Study each of them with other variables like gender, sector, purpose of loan, and create a table to brainstorm and see correlation with the other data. • Do univariate analysis, frequency (mainly categorical) and distribution analysis (mainly continuous), then proceed to bivariate analysis, see if the data is normally or non-normally distributed. Use different test methods accordingly. If both categorical, use contingency analysis/cross tabulation and kai

		square test to test if statistically significant. From observation, create hypothesis and test it. Correlation is only if both are continuous (Pearson). Study seasonal patterns, whether people tend to borrow more towards year-end, identify the reasons for it (such as culture, tradition, weather if patty field crops die during monsoon). Do text mining at the later phase. Do not compare at a global level – even at a country level, there is much variation. Philippines is a predominantly Christian country but the southern part is mostly made up of Muslims. • Focus and decide on specific countries, as narrowing down provides more solid analysis. Philippines has income and expenditure data up to the region level.
2.	Follow-up actions	• Create data dictionary for all the data variables in all files, including the description and data type (numerical, categorical) and conduct basic exploratory analysis.

Item Due (Team) / Actions
Deadline: End of week. 1. Data dictionary of all the terms, and begin exploratory data analysis for sharing with sponsor next meeting.