ANLY482 Internal MEETING MINUTES (3 Mar 2018)

Date:	3 Mar 2018
Time:	11:00 – 17:30
Venue:	Library Project Room 4.7
Attendees:	Team: Ruiyan, Qian, Nicholas
Agenda:	1. Identify the data issues of each file 2. Discuss how to do data cleaning and preparation 3. Perform data cleaning and preparation

S/N	Things Discussed/Done	Remark
1	kiva_loans.csv	Data issues: • 'region': some data has encoding problem. Most of the regions from Vietnam have number, such as 01, 02, …, 10 in front of the region name. Most of the regions contain province, city, and county level. • Loan id '1281091' has funded time before posted time, which is impossible. • 'tags': may contain multiple tags separated by comma. • 'borrowers_gender': may contain multiple genders. • 'date': the date that the loan is posted, which is duplicated with 'posted_time'. • 'currency': not important since 'loan_amount' and 'funded_amount' are in USD. Way of data cleaning and preparation: • 'region': plan to separate region into different levels, such as province, city and country by referring to the additional dataset from gdam.org. We will perform this data preparation after deciding which countries to focus on. • 'currency': we decide to remove this column because 'funded_amount' &

		‘loan_amount’ is in USD. Therefore, it is irrelevant for the analysis. • Remove record with loan id = ‘1281091’. • ‘tags’: create column for each unique tag containing binary value. For example, the new column ‘#Animals’, if value is 1, then indicates this loan has this tag. • ‘borrowers_gender’: create 2 columns ‘female_count’, ‘male_count’ and ‘number of borrowers’, which contain the number of female and male borrowers for this loan. • Remove ‘date’. • Remove ‘currency’.
2	kiva_mpi_region _locations.csv	Data issues: • 1788 records are missing all variables, except for ‘geo’, which is also invalid values (1000, 1000). Way of data cleaning and preparation: • Delete those 1788 records.
3	loan_theme_ids.csv	Data issues: • 14813 records are missing ‘Loan Theme ID’, ‘Loan Theme Type’ and ‘Partner ID’. Way of data cleaning and preparation: • Remove those records.
4	loan_themes_by_region.csv	Data issues: • ‘geocode_old’: the Kiva’s old geocoding system and contains lots of missing values. • ‘geocode’ & ‘geo’: lon-lat pair, which is deplicated information with ‘lon’ and ‘lat’. Way of data cleaning and preparation: • Remove ‘geocode_old’. • Remove ‘geocode’. • Remove ‘geo’.

5	loans.csv (additional data)	Inner join loans.csv and kiva_loans by id. Data issues: • 671204 out of 1419607 records matched. Way of data cleaning and preparation: • Remove those non-matching records. • Remove those columns in loans.csv that are duplicated with kiva_loans.csv: o ‘loan_amount’, ‘funded_amount’ o ‘activity_name’, ‘sector_name’ o ‘loan_use’, ‘country_code’ o ‘country_name’, ‘town_name’ o ‘currency’, ‘partner_id’ o ‘posted_time’, ‘disburse_time’ o ‘raised_time’, ‘tags’ o ‘borrower_genders’, ‘repayment_interval’ • Remove those columns in loans.csv that we feel are more important: o ‘loan_name’, ‘lender_term’, o ‘num_lenders_total’ o ‘num_journal_entries’ o ‘num_bulk_entries’ o ‘borrower_pictured’ o ‘currency_policy’ o ‘currency_exchange_coverage_rate’
6	Data preparation	• Create ‘rate of funding’: funded amount/ time difference between posted time and funded time (in days) • Create ‘average loan amount’: loan amount/ number of borrowers • Create ‘monthly repayment per borrower’: funded amount/ number of borrowers/ terms in month

Item Due (Team) / Actions

Deadline: Mar 6

1. Qian:
   - Study the distribution of loan amount and number of loans breakdown by other factors, such as gender, sector, country, region, activity (purpose of loan) and loan theme type.
   - To find out the countries or areas where Kiva has the most active loans.
2. Ruiyan:
   - Study the distribution of repayment period (terms in month) breakdown by other factors such as gender, sector, country, region, activity (purpose of loan) and loan theme type.
   - Study the distribution of repayment ability (loan amount/ terms in month) breakdown by other factors, such as countries, regions, gender, sector, purpose of loan, and loan theme type.
3. Nicholas:
   - Study the distribution of rate of funding (funded amount/time difference between posted time and funded time) breakdown by other factors, such as countries, regions, gender, sector, purpose of loan, and loan theme type.
4. All:
   - To find out the borrowing patterns (in terms of purpose of loan, loan amount, number of loans, loan theme type and sector) of different countries, regions and gender.
   - Complete individual’s meeting minutes and upload to Google drive.