# PROPOSAL DOCUMENT

GIT URL for project: https://github.com/tuli0018/KIVA-data-analysis

## Data Overview:

I am planning on using KIVA open dataset which is in CSV format. Kiva.org online is a crowdfunding platform to extend financial services to poor and financially excluded people around the world. Kiva is a 501 non-profit organization that allows people to lend money via the Internet to low-income entrepreneurs and students in 77 countries. Kiva lenders have provided over 1 Billion USD of loan to over 2 million people. I am planning on primarily using 2 KIVA datasets, one is the total number of loans that are being funded at the moment along with the loan requester data such as amount, total funding achieved, total funding pending etc. This dataset is 3394 rows of data. The second file is a list of lender data from across the world. This data lists each lender name, location, their primary interest for lending, location information and the different loan themes that the lender promotes. The second file has 21 columns and 1531 rows of data. I have found another dataset that has a list of locations that with their MPI index (multidimensional poverty index) that can potentially be used for cross sectional analysis. All 3 files are in CSV format.

## Here is a list of interesting statistical questions that I plan to answer using Spark API:

1) Which country and region has the highest poverty?

A. I used the KIVA MPI dataset to solve this problem. First, we get the MPI value using the sort by operation on the MPI data RDD. This MPI value is a decimal point value that is then used to find the country, local and world regions respectively using the filter operation on RDDs.

2) Total # and amount of loan requested by the country of highest poverty.

A. This method takes in 2 parameters, the loans data frame and the MPI data RDD. To answer the analytical question, first I parsed out problem 1's result to input in the second problem. Then using filter operation on the loans data frame, I found the total # of loans the amount requested by the country with the highest poverty.

3) List of lenders that can fulfill this loan.

A. This method takes in 2 data frames, lenders and loans. I created temporary views to be able to use spark SQL to query my sub results and then filter down using count and ORDER BY operations in SQL. To answer this analytical question, I wanted to provide a list of 2 lenders that are can fund the loans. First recommendation was based on the country and theme rankings. Second recommendation was on a global scale, just based on theme rankings.

4) Which lender is most likely to fund their loan?

A. This was fairly easy to analyze. I took the 2 lenders and its corresponding data to perform a compare and contrast of all their rankings.

5. Which sector of the market has the highest requirement for loans?

A. This problem was easily solved using spark SQL on loans dataset that was called into the method as a data frame.

6. Total # of lenders that specialize loans in that sector of the market?

A. Here, I called the previous method to get the results and parse them out. The result was then used to query the lenders data frame using spark SQL API to get the total number of vendors that specialized in the market sector that needed the most amount of loans to be funded.

7. Information (loan ID, requester name, partner name, loan amount etc.) of the loan that is most likely to be fundraised.

A. To answer this question, I had to query the loans data frame to get the lowest amount needed value, which would mean that the loans have the highest probability of getting fulfilled. Then, using this value, I queried all the loans that had the amount needed = lowest value. Once I had that, I compared each of the 2 loans to predict which one was going to get fulfilled first.

8. Which sector has the least chance of loans to be fundraised.

A. To answer this question, I found the business sector where the loans with the lowest average borrower ratings originated from.

9. Which country has the highest # of loans that are most likely to be fundraised.

A. To answer this question, firstly, I got a list of top 3 countries that have the highest average borrower ratings irrespective of the total number of loans requested from that country. Then, for each of these countries, I found the country that has the least amount of loans that are requested and that was the country that would have all their loans fundraised due to the amount of loans and the average borrower's ratings.

I plan on using Spark Scala API (Scaladocs). The reason I want to use Spark Scala API is because I want to be able to split my compute between datasets and dataframes. Spark Scala API has good documentation, rich semantics and several ready to use operations that perform complex calculations such as map, filter, groupBy by simply specifying the task and column/row details. It is also easy to filter the data using these APIs.

My strong forte is in Java. I want to enhance my skills in Scala.

**Surprising finds**

The most surprising find I had from analyzing this dataset was that Education was the sector that had the least or the lowest chance of getting fundraised, from a theme perspective.

GIT URL for project: https://github.com/tuli0018/KIVA-data-analysis