Exploring Kiva Loans Data

Gabriel Majivu

11/29/2016

1. About

There has been a shift from traditional credit sources like banks to microlending institutions for individuals and small businesses who find it hard to access bank loan facilities. Microlending organizations have taken to cater for this niche market.

This project has the following goals:

1. To predict if a loan application will attract funding in full, or partially or none at all. This will help borrowers to optimize their loan applications.
2. To help field partners discover insights that will help them understand their customers better and create a good customer experience for them by advising on their loan applications.

In this project we use data from Kiva. Kiva is an international non-profit that allows people to lend money to low-income entrepreneurs and students in over 80 countries. Its mission is to enable people to create opportunity for themselves and for others. In brief, Kiva supports microlending across the whole world. They offer loans either directly to borrowers (only in the US) or through appointed field partners who assess the loans need and administer the loans on behalf of the lenders.

Based on this project findings, a field partner will be able to leverage on the insights derived from this study to informatively advise the borrowers who seek funds for their projects.

2. Data

Kiva exposes an API for getting their data. I wrote a script to facilitate getting the data.

```
load('./loans.RData')
```

2.1 Inspecting the Loans Dataset

Data types for columns in the loan dataset:

```
str(loans, vec.len=2)
```

```
## 'data.frame':    1176398 obs. of  31 variables:
##  $ id                           : int  1200150 1200155 1200483 1200488 1200493 ...
##  $ name                         : chr  "Pamela" "Mercy" ...
##  $ status                       : chr  "fundraising" "fundraising" ...
##  $ funded_amount                : int  0 0 0 0 0 ...
##  $ basket_amount                : int  0 0 0 175 0 ...
##  $ activity                     : chr  "Home Energy" "Home Appliances" ...
##  $ sector                       : chr  "Personal Use" "Personal Use" ...
##  $ themes                       : chr  "Green" "Green" ...
##  $ use                          : chr  "to buy a solar lantern." "to buy an eco-friendly stove." ...
##  $ partner_id                   : int  156 156 145 145 145 ...
##  $ posted_date                  : chr  "2016-12-13T17:40:06Z" "2016-12-13T17:40:06Z" ...
##  $ planned_expiration_date      : chr  "2017-01-12T17:40:06Z" "2017-01-12T17:40:06Z" ...
##  $ loan_amount                  : int  75 50 225 175 200 ...
##  $ borrower_count               : int  1 1 1 1 1 ...
```

```
##  $ lender_count                 : int  0 0 0 0 0 ...
##  $ bonus_credit_eligibility     : logi  FALSE FALSE TRUE ...
##  $ tags                         : chr  "#Eco-friendly, #Technology" NA ...
##  $ description.languages        : chr  "en" "en" ...
##  $ image.id                     : int  2384072 2384077 2384595 2384593 2384599 ...
##  $ image.template_id            : int  1 1 1 1 1 ...
##  $ location.country_code        : chr  "KE" "KE" ...
##  $ location.country             : chr  "Kenya" "Kenya" ...
##  $ location.town                : chr  "Kitale" "Kitale" ...
##  $ location.geo.level           : chr  "town" "town" ...
##  $ location.geo.pairs           : chr  "1.016667 35" "1.016667 35" ...
##  $ location.geo.type            : chr  "point" "point" ...
##  $ currency_exchange_loss_amount: num  NA NA NA NA NA ...
##  $ video.id                     : int  NA NA NA NA NA ...
##  $ video.youtubeId              : chr  NA NA ...
##  $ video.title                  : chr  NA NA ...
##  $ video.thumbnailImageId       : int  NA NA NA NA NA ...
```

An overview of the data variables is below:

```
names(loans)
```

```
##  [1] "id"                           "name"
##  [3] "status"                       "funded_amount"
##  [5] "basket_amount"                "activity"
##  [7] "sector"                       "themes"
##  [9] "use"                          "partner_id"
## [11] "posted_date"                  "planned_expiration_date"
## [13] "loan_amount"                  "borrower_count"
## [15] "lender_count"                 "bonus_credit_eligibility"
## [17] "tags"                         "description.languages"
## [19] "image.id"                     "image.template_id"
## [21] "location.country_code"        "location.country"
## [23] "location.town"                "location.geo.level"
## [25] "location.geo.pairs"           "location.geo.type"
## [27] "currency_exchange_loss_amount" "video.id"
## [29] "video.youtubeId"              "video.title"
## [31] "video.thumbnailImageId"
```

1. **id**: unique id of the borrower
2. **name**: name of the borrower
3. **status**: current funding status that a loan has. Options include:
    - fundraising - The loan has not yet been funded. You can find the amount funded so far by checking *funded_amount*.
    - funded - This loan request has been completely funded and is not available for new loans by lenders.
    - in_repayment - The loan has been disbursed to the borrowers and they are in the process of using the funds and making payments on the loan to the field partner.
    - paid - The loan has been paid back in full by the borrower.
    - defaulted - A loan which has remained delinquent 6 months after the end of the loan payment schedule.
    - refunded - Refund the funded portion of the loan to lenders after the loan has been partially funded, fully funded, or even during repayment.
4. **funded_amount**: This is the amount of the loan which has been purchased by Kiva lenders.
5. **basket_amount**: This is the amount of the loan which lenders have saved in their shopping baskets, but has not been confirmed as purchased.

6. **activity**: Type of activity the borrower is involved with.
7. **sector**: Type of sector the borrower is involved in.
8. **themes**: General themes that further categorise borrowers.
9. **use**: Description for the purpose of the loan.
10. **partner_id**: Unique id for the field partners.
11. **posted_date**: The date the loan was posted on the Kiva website.
12. **planned_expiration_date**: The date the loan bid expires on the Kiva website. It is not set for some loans.
13. **loan_amount**: Amount of funding sought by the borrowers in a bid for a loan.
14. **borrower_count**: The number of borrowers for a loan. It is a good indication for single vs group borrowers.
15. **lender_count**: Number of lenders who have purchased a loan.
16. **bonus_credit_eligibility**: Whether a loan application is eligible for extra credit.
17. **tags**: Borrower chosen tags to classify their loan applications.
18. **description.languages**: Language choice(s) for the borrower.
19. **image.id**: Image ID for a borrower.
20. **image.template_id**: Image template id.
21. **location.country_code**: Country code of borrower.
22. **location.country**: Country name of borrower.
23. **location.town** Town of the borrower.
24. **location.geo.level**: Level of accuracy of supplied geometry.
25. **location.geo.pairs**: The coordinate pairs for the geometry.
26. **location.geo.type**: The type of geometry defined by the coordinate pairs provided.
27. **currency_exchange_loss_amount**: Losses realized by the lender due to fluctuations in the value of the local currency against the US dollar. This will result in the paid amount of loan being less than the full loan amount even when the status of the loan is listed as *paid*.
28. **video.id**: Id for video.
29. **video.youtubeId**: Id for video on youtube.
30. **video.title**: Title of the video.
31. **video.thumbnailImageId**: Thumbnail id for the video.

The loans dataframe has over 1.1M rows of observations.

```
dim(loans)
```

```
## [1] 1176398      31
```

Summary statistics of the dataset:

Check the number of null values (*NAs*) per variable

```
character_cols <- names(loans)[sapply(loans, is.character)]
for(col in character_cols){
  na_strings <- sum(loans[[col]]=="NA")
  nas <- sum(is.na(loans[[col]]))
  print(paste(col, ": ", na_strings,',NA:' ,nas))
}
```

```
## [1] "name :  0 ,NA: 0"
## [1] "status :  0 ,NA: 0"
## [1] "activity :  0 ,NA: 0"
## [1] "sector :  0 ,NA: 0"
## [1] "themes :  NA ,NA: 871811"
## [1] "use :  0 ,NA: 0"
## [1] "posted_date :  0 ,NA: 0"
## [1] "planned_expiration_date :  NA ,NA: 369087"
## [1] "tags :  NA ,NA: 749871"
```

```
## [1] "description.languages :  0 ,NA: 0"
## [1] "location.country_code :  9 ,NA: 0"
## [1] "location.country :  0 ,NA: 0"
## [1] "location.town :  NA ,NA: 130659"
## [1] "location.geo.level :  0 ,NA: 0"
## [1] "location.geo.pairs :  0 ,NA: 0"
## [1] "location.geo.type :  0 ,NA: 0"
## [1] "video.youtubeId :  NA ,NA: 1175864"
## [1] "video.title :  NA ,NA: 1175878"
```

Check the proportion of null values (*NAs*) per variable

2.2 Kiva by numbers

A brief look into the data set highlighting important numbers:

- The number of countries Kiva has facilitated loans:

```
length(unique(loans$location.country_code))
```

```
## [1] 91
```

- The number of unique borrowers over time:

```
length(unique(loans$id))
```

```
## [1] 1173553
```

- The total amount of loans funded:

```
sum(loans[['funded_amount']])
```

```
## [1] 949203295
```

- Loan repayment rate:

```
sum(loans[['funded_amount']]) / sum(loans[['loan_amount']]) * 100
```

```
## [1] 96.37544
```

2.3 Data munging

Remove columns that we may not need which include: *id*, *image.id*, *image.template_id*, *video.id*, *video.youtubeId*, *video.title*, *video.thumbnailImageId*

```
loans$id <- NULL
loans$image.id <- NULL
loans$image.template_id <- NULL
loans$video.id <- NULL
loans$video.youtubeId <- NULL
loans$video.title <- NULL
loans$video.thumbnailImageId <- NULL
```

Each loan has a status at any given time. We set these loan statuses as factors:

```
loans$status <- factor(loans$status,
                       levels = unique(loans$status),
                       labels = c("Fundraising", "Funded", "Expired"))
table(loans$status)
```

```
##
## Fundraising      Funded     Expired
##        7247     1123882       45269
```

Similarly, we set loan sector as factors:

```
loans$sector <- factor(loans$sector)
table(loans$sector)
```

```
##
##    Agriculture           Arts       Clothing   Construction      Education
##         281033          22493          70000          16718          34457
##  Entertainment           Food         Health        Housing  Manufacturing
##           1751         274649          12380          48160          13944
##   Personal Use         Retail       Services Transportation      Wholesale
##          32625         246537          85848          33910           1893
```

We also set latitude and longitude from *location.geo.pairs*

```
loans <- separate(loans, location.geo.pairs,
                  into=c('location.geo.lat', 'location.geo.long'),
                  sep=' ',
                  remove = TRUE)
loans$location.geo.lat <- as.numeric(loans$location.geo.lat)
loans$location.geo.long <- as.numeric(loans$location.geo.long)
```

Most loans are administered through Kiva field partners. However, some loans are given directly to the borrowers. These are identified as *direct* loans. We add a column to identify direct vs non-direct loans. Direct loans do not require field partners hence partner_id is null.

```
loans$direct <- as.numeric(is.na(loans$partner_id))
```

Transform the date columns.

```
loans$posted_date <- as.Date(loans$posted_date)
loans$planned_expiration_date <- as.Date(loans$planned_expiration_date)
```

2.4 Data Exploration

This section aims to gain insights for some variables of interest.

2.4.1 status

status implies the stage a loan has. There are a number of stages as outlined in the Kiva website. Loans for this dataset have the following status:

Here is the distribution by status of the loan.

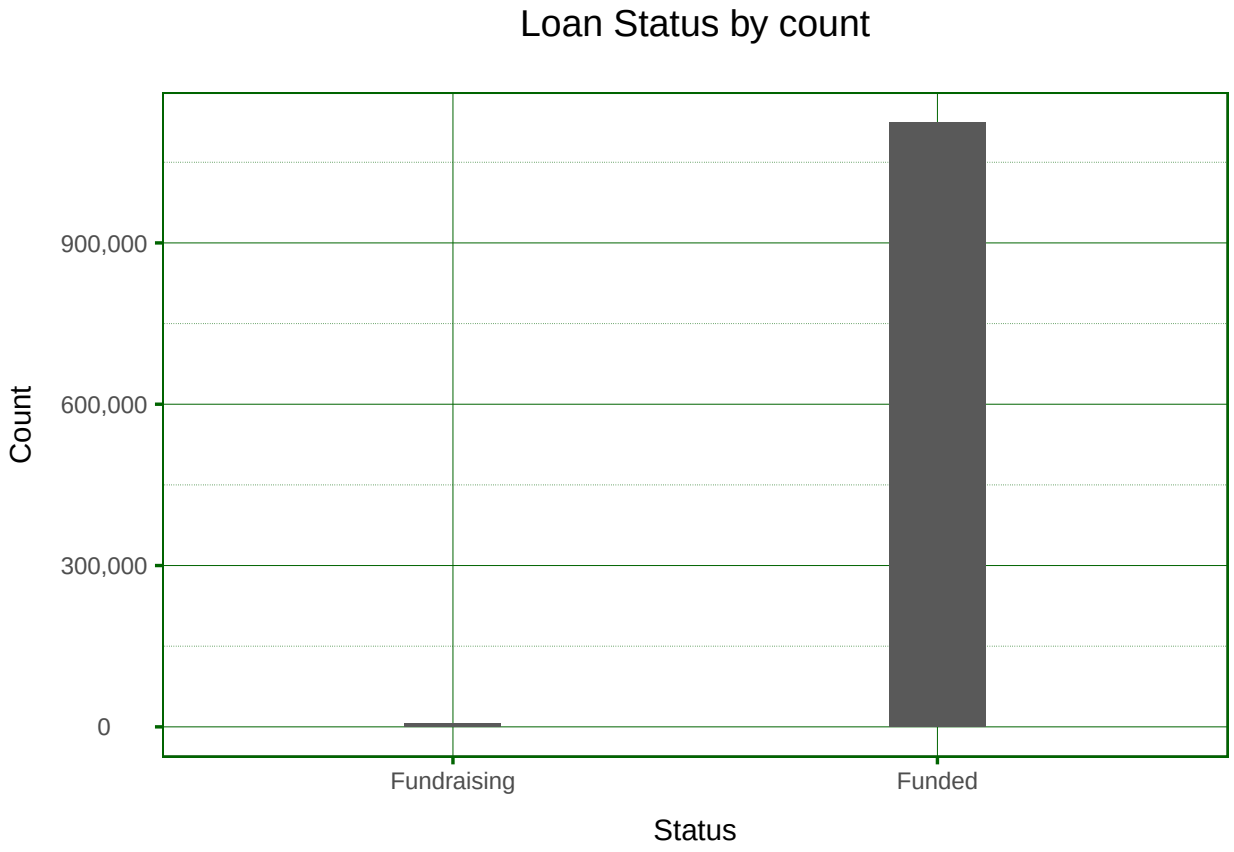
```
status_dist <- data.frame(table(loans$status))
status_dist$Percentage <- status_dist$Freq / sum(status_dist$Freq) * 100
colnames(status_dist) <- c('Status', 'Frequency', 'Percentage')
status_dist
```

```
##        Status Frequency Percentage
## 1 Fundraising      7247   0.616033
## 2      Funded   1123882  95.535865
## 3     Expired     45269   3.848102
```

95.54% of the loans have been funded and only 0.62% are in the fundraising stage. Unfortunately, the "Expired" status has not been explained on the kiva website so it is ambiguous. For these reasons, we shall remove rows with this status from the dataset.

```
loans <- subset(loans, status=='Fundraising' | status=='Funded')
```

A plot for the status variable:

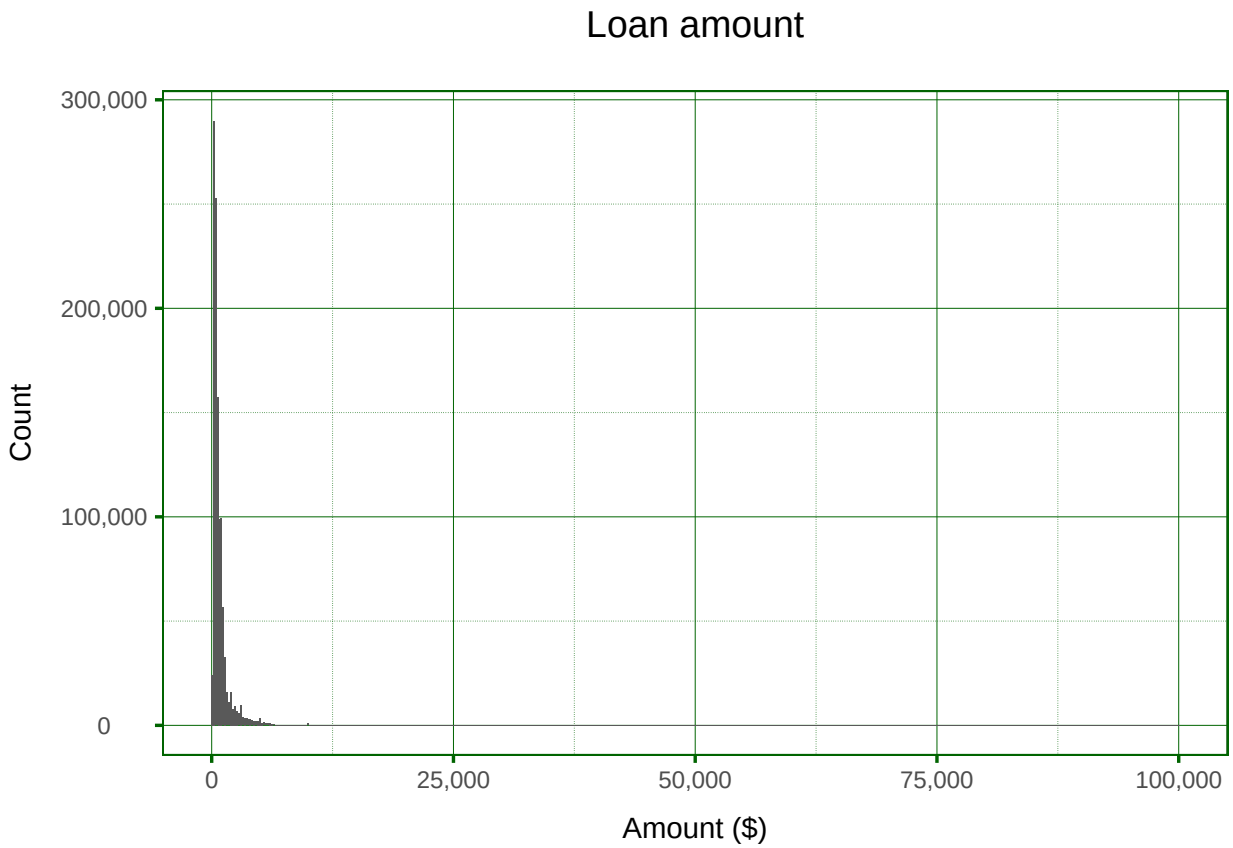


2.4.2 loan_amount

```
summary(loans$loan_amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      25     300     500     813    1000  100000
```

The average amount of loan borrowed is $813. We also observe that the median is lower than the mean by a few hundreds of dollars. This points to existence of high loan amounts. The lowest amount requested is $25, the highest being $10,000.



There are a number high loan amounts that cause skew on the plot. Since the focus on this study is to evaluate microloans, we need to compute a sensible cutoff to remove the high loan amounts.

```
# choosing a number of cutoff amounts
cut_off_amounts <- c(5000, 6000, 7000, 8000, 9000, 10000, 11000, 12000)
for(amount in cut_off_amounts){
    count <- sum(loans$loan_amount > amount)
    proportion <- 100 * count / dim(loans)[1]
    print(paste("cut_off_amount =", amount,
                ", count =", count,', proportion =' ,proportion, '%'))
}
```

```
## [1] "cut_off_amount = 5000 , count = 9951 , proportion = 0.87974050705092 %"
## [1] "cut_off_amount = 6000 , count = 4561 , proportion = 0.403225449970781 %"
## [1] "cut_off_amount = 7000 , count = 2473 , proportion = 0.21863111988111 %"
## [1] "cut_off_amount = 8000 , count = 1766 , proportion = 0.156127196809559 %"
## [1] "cut_off_amount = 9000 , count = 1391 , proportion = 0.122974479480236 %"
## [1] "cut_off_amount = 10000 , count = 198 , proportion = 0.0175046347498826 %"
## [1] "cut_off_amount = 11000 , count = 182 , proportion = 0.0160901188104982 %"
## [1] "cut_off_amount = 12000 , count = 155 , proportion = 0.0137031231627869 %"
```

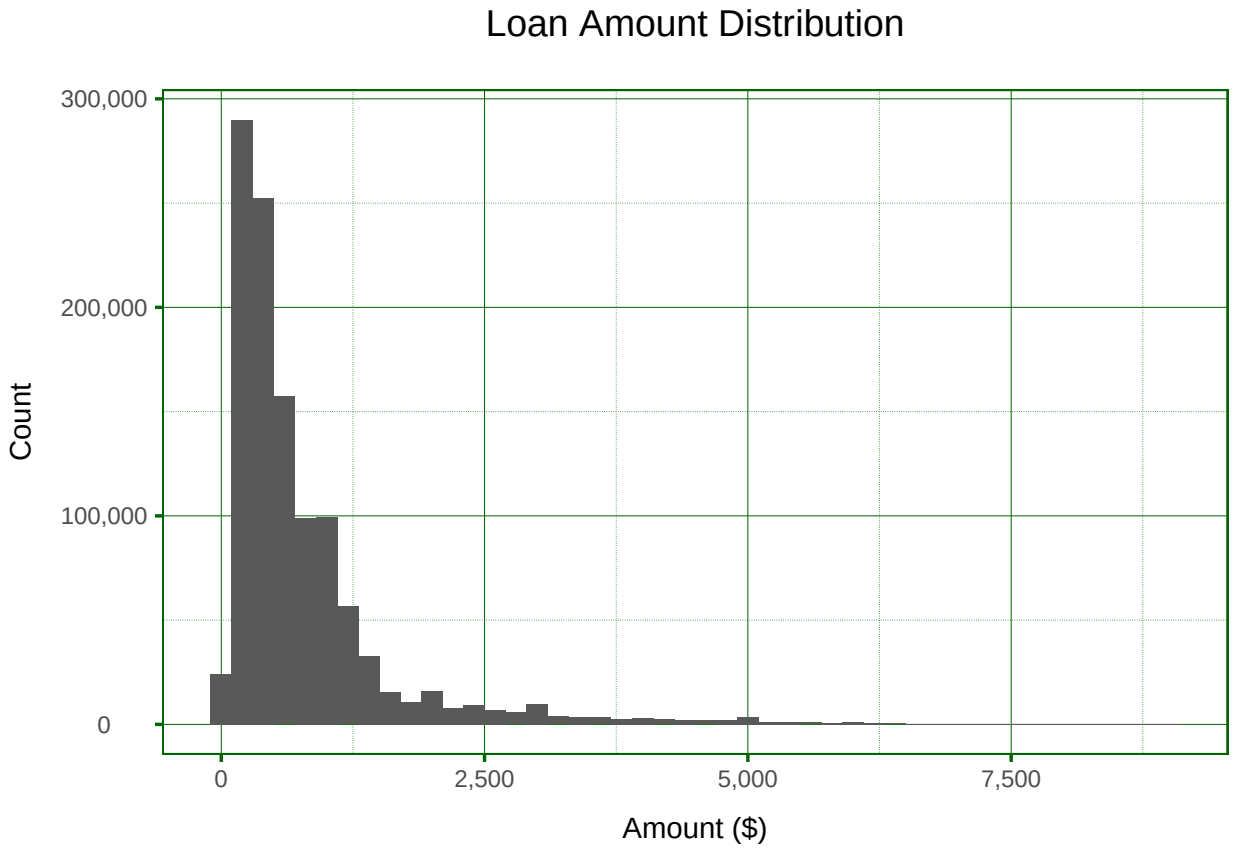
For a cutoff at an amount of $9,000 only 1,391 observations will be removed. This represents about 0.13% of data. Further, $9,000 is practically a large sum and any amount above it may be safely ignored for purposes of our study.

```
loans <- subset(loans, loans$loan_amount <= 9000)
summary(loans$loan_amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##    25.0   300.0   500.0   799.5  1000.0  9000.0
```

With the new dataset, 75% of the loan amounts borrowed are less or equal to $1000 and half the loans are not more than $500. The average amount of loan borrowed is $799.50.



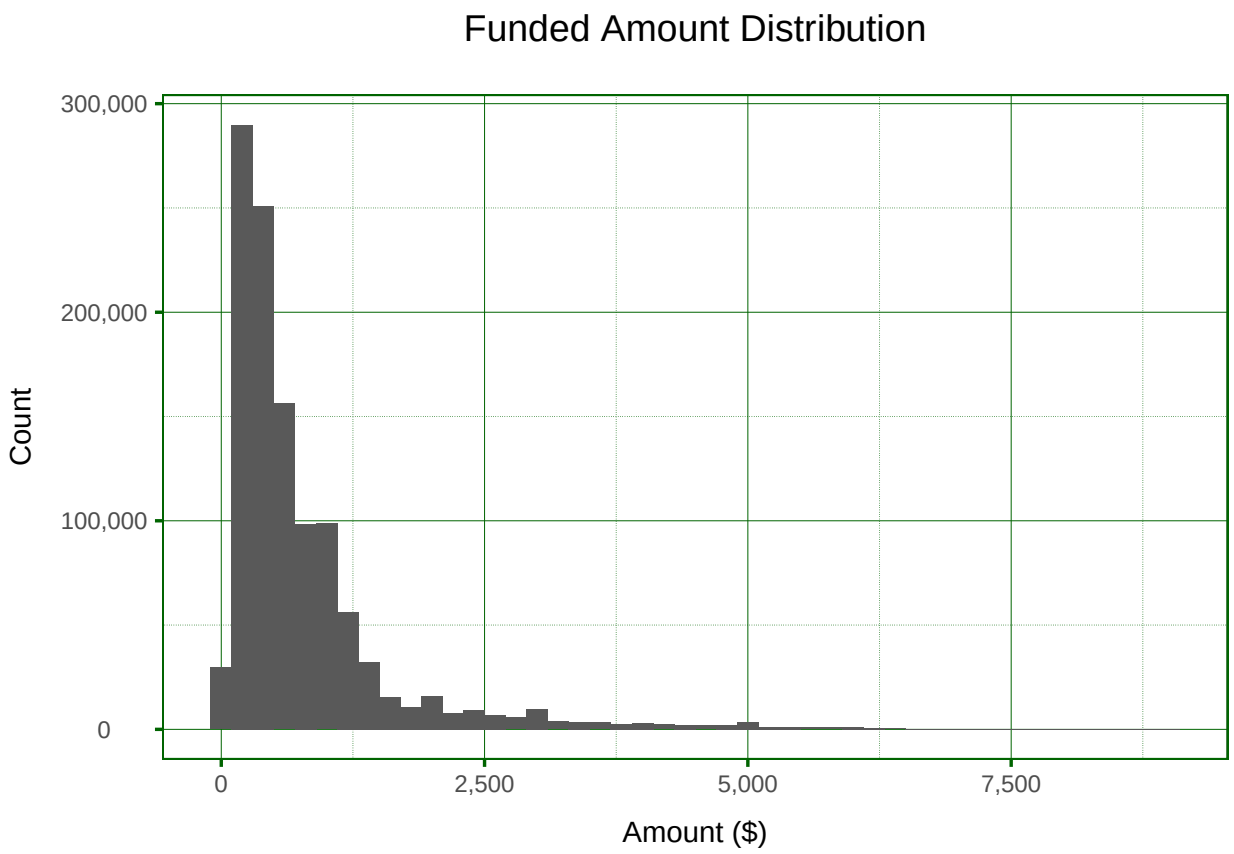
Here, it is much clearer that most borrowers go for less than $1,250 (3rd quartile amount is $1,000).

2.4.3 funded_amount

```
summary(loans$funded_amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   300.0   500.0   794.2   975.0  9000.0
```

The average funded amount for a loan is $794.20. Some loans have $0 funding - these are probably loans under fundraising. We have a naive funding success rate of 96.5% by comparing with the average loan amount.

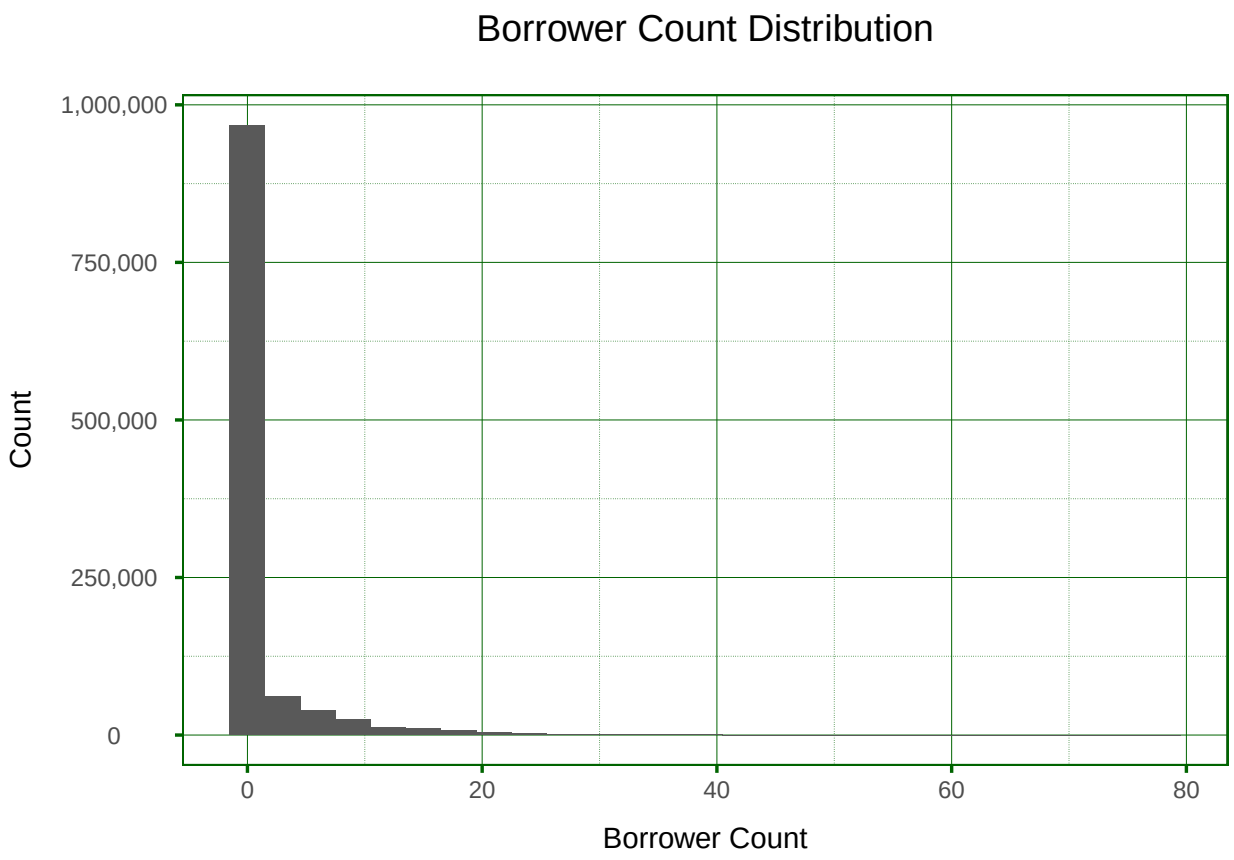


The plot shows a similar distribution to that of *loan_amount*. The *funded_amount* is a proportion of the *loan_amount* as bid by the lenders. Perhaps there could be cases where a loan is oversubscribed. We shall see if this is the case in further analysis.

2.4.4 borrower_count

```
summary(loans$borrower_count)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.974   1.000  79.000
```

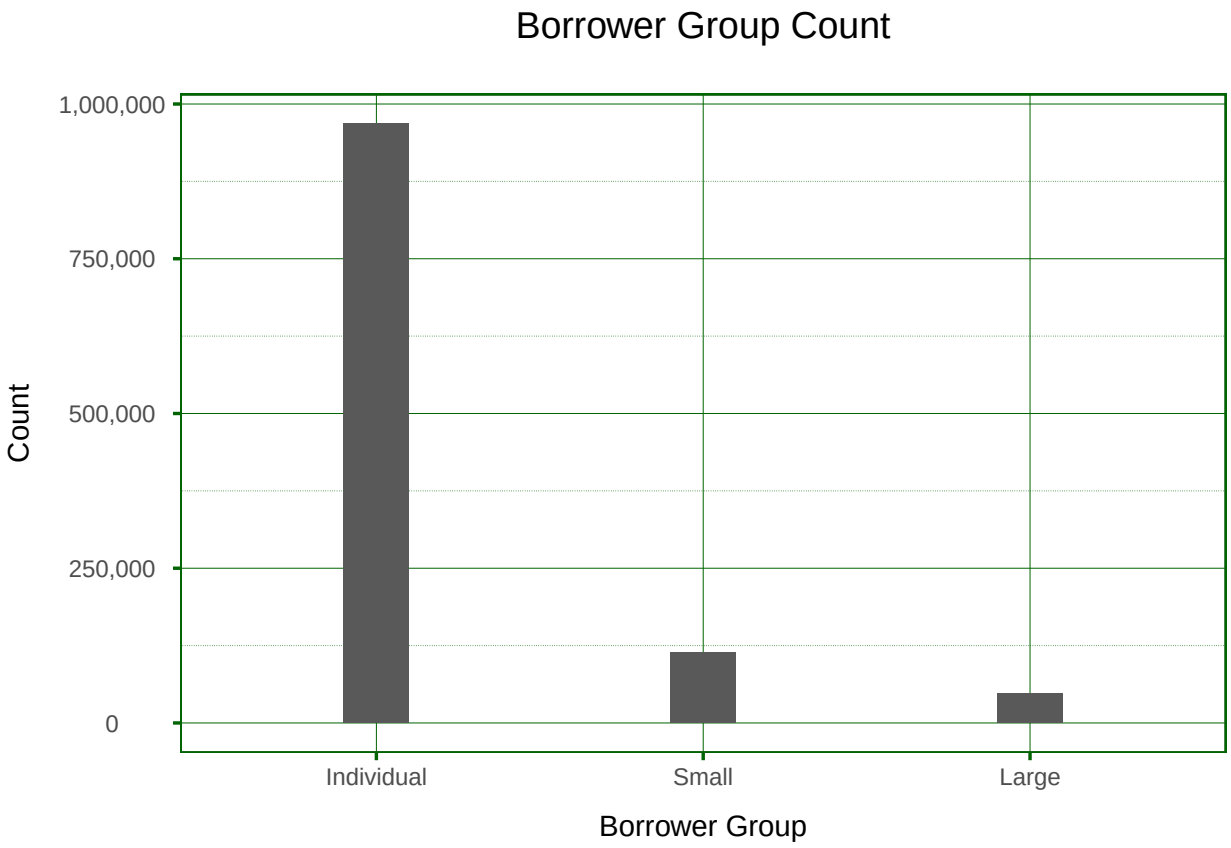


Most loans have individual borrowers or small groups i.e. few people borrowing together. The *borrower_count* might be more valuable if it is transformed to categories with specific number of people. Making groups for this variable thus: * individual = 1 persons * small group = 2 - 9 persons * large group = more than 10 persons

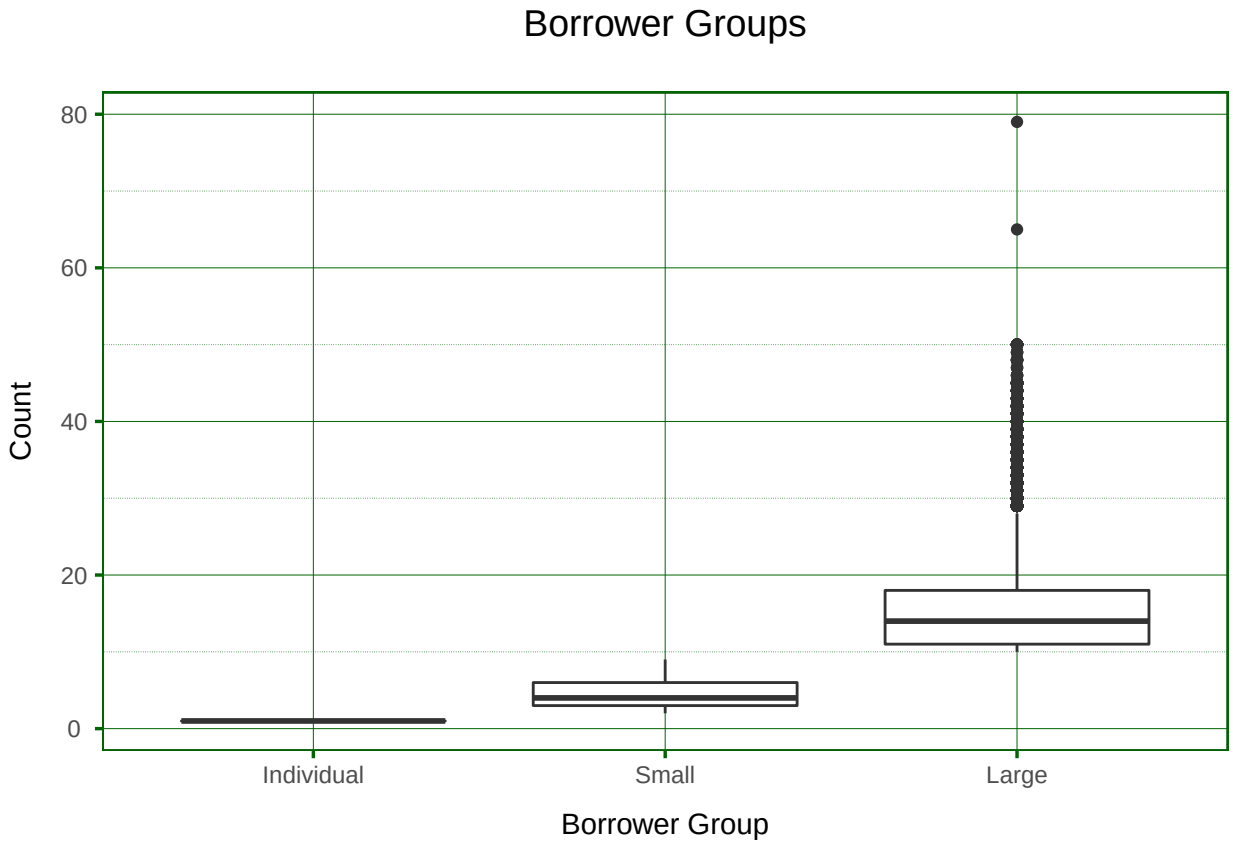
We therefore create a new variable *borrower_group*

```
loans$borrower_group <- cut(loans$borrower_count,
                            c(0,1,9,79),
                            labels=c('Individual','Small','Large'),
                            ordered_result = TRUE)
```

Plotting the new *borrower_group* variable:



How does a box plot of the same variable look like?



The "Large" group shows the greatest variability. It has considerable outliers but since the overall count in this group is small (as seen from the preceding plot) it is not necessary to break it down further. The average for each group is 1, 5, and 14 respectively.

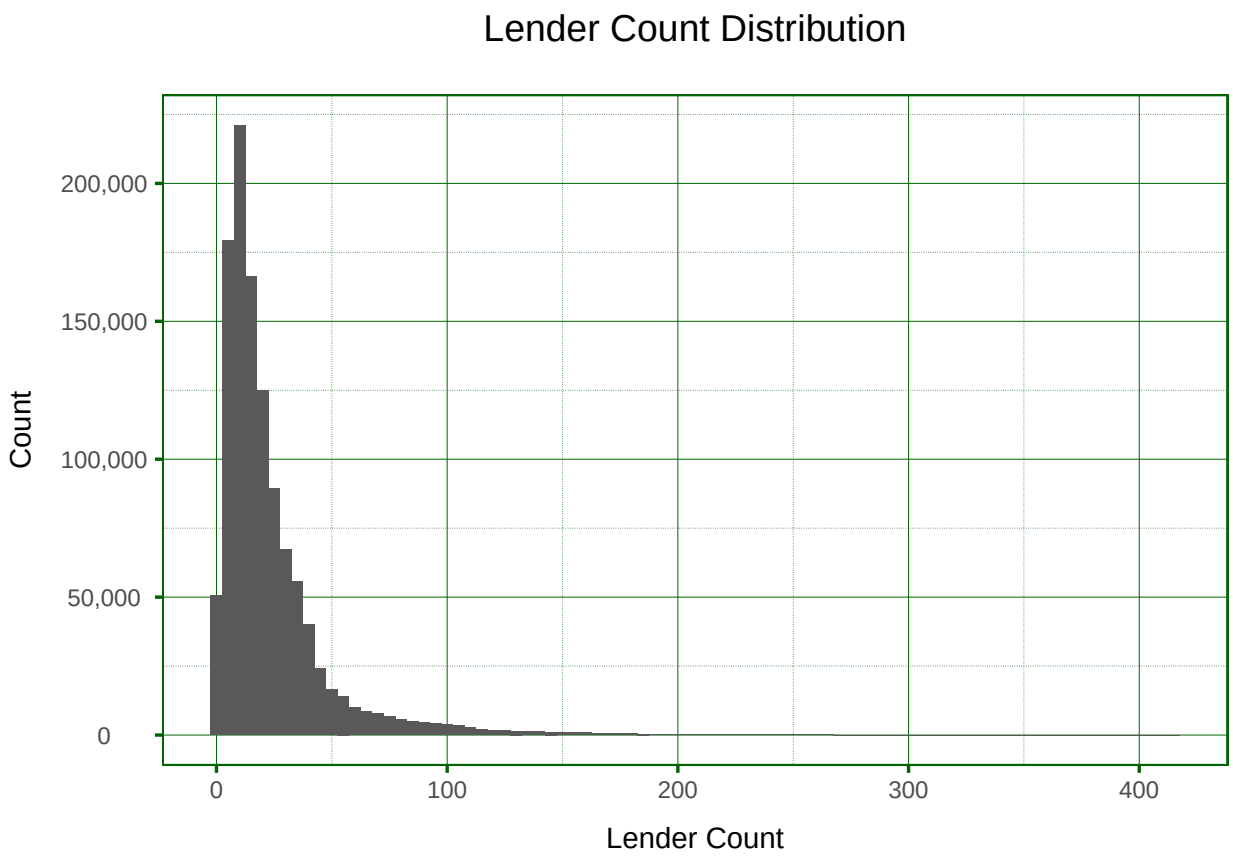
2.4.5 lender_count

This variable probably also needs grouping. We could apply the same treatment on this group as the *borrower_count*.

```
summary(loans$lender_count)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   16.00   23.15   29.00  413.00
```

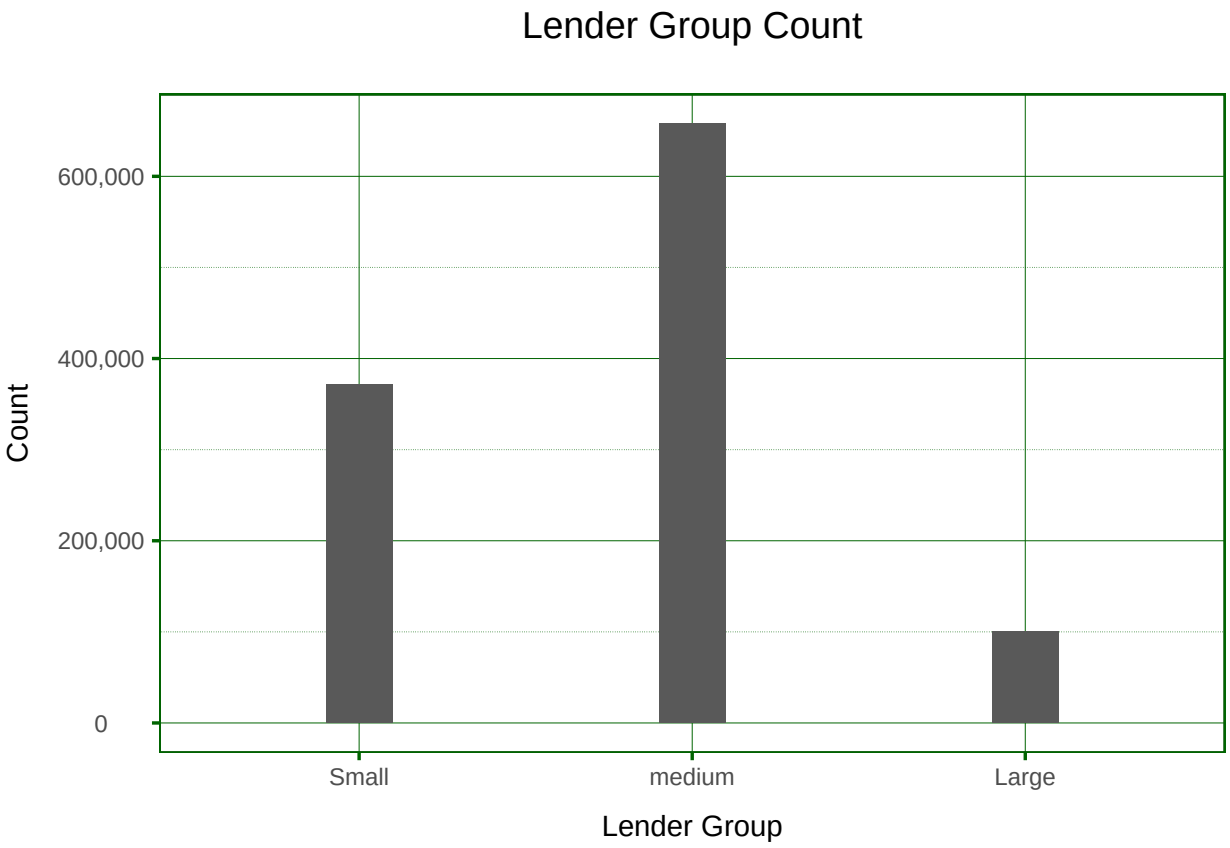
Lender group average 23 persons per loan. Lenders tend to spread their risk by grouping and buying into loans in small amounts



From the plot we see most loans having fewer than 50 lenders. We can create 3 lender groups thus: * small group = less than 10 persons * medium group = 10 - 50 persons * large group = more than 50 persons

```
loans$lender_group <- cut(loans$lender_count, c(0,10,50,2986),
                          labels=c('Small', 'medium','Large'),
                          ordered_result = TRUE,
                          include.lowest = TRUE)
```

Plotting the new *borrower_group* variable:



From the plot we see most loans belong to the medium group (10 - 50 lenders). Loans which attract more than 100 lenders are much fewer than the other two categories combined. THis is also exlplained by the fact that loans with high amounts are much less and it is these loans that would require many more lenders to combine effort - perhaps to minimize their risk.

2.4.6 sector

Borrowers have to fill in the sector in which they operate. This variable may be used to assess the purpose of funding. The sectors in this dataset are:

```
levels(loans$sector)
```

```
##  [1] "Agriculture"    "Arts"           "Clothing"       "Construction"
##  [5] "Education"      "Entertainment"  "Food"           "Health"
##  [9] "Housing"        "Manufacturing"  "Personal Use"   "Retail"
## [13] "Services"       "Transportation" "Wholesale"
```

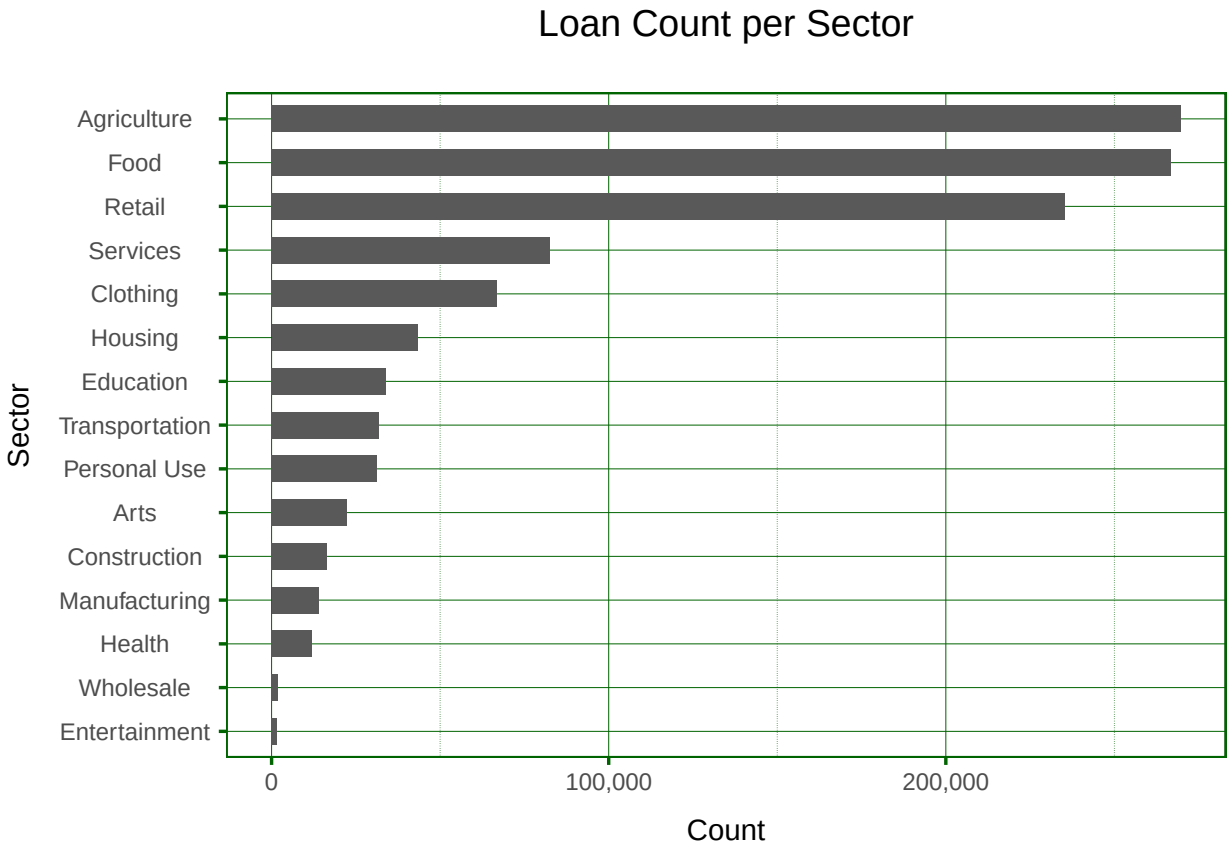
By count and proportion:

```
table(loans$sector)
```

```
##
##    Agriculture           Arts       Clothing   Construction      Education
##         269654          22421          66787          16386          33980
##  Entertainment           Food         Health        Housing  Manufacturing
##           1717         266836          11938          43298          13928
##   Personal Use         Retail       Services Transportation      Wholesale
##          31242         235295          82462          31947           1847
```

```
sector_dist <- data.frame(sort(table(loans$sector)))
sector_dist$Percentage <- sector_dist$Freq / sum(sector_dist$Freq) * 100
colnames(sector_dist) <- c('Sector', 'Frequency', 'Percentage')
# sector_dist
```

A plot to visualize the information above:



Evidently, the most active sectors include "agriculture", "food", "retail". It seems accessing food is the most important activity on the basis of "agriculture"" and "food" sectors topping by the highest number of loans. Another observation is that small scale business ventures like "retail"" and "services" rank higher than the more capital intensive ventures like "construction", "manufacturing" and "wholesale". "entertainment" is the least popular sector - no one wants to fund leisure activities.

2.4.7 funding_rate

This is a new variable which defines the rate of success in funding bid by the borrower, in other words how much of the original sought amount was actually funded?

```
loans$funding_rate <- 100*loans$funded_amount / loans$loan_amount
summary(loans$funding_rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00  100.00  100.00   99.43  100.00  100.00
```

The average rate of funding is 99.43%. Funding rate is considered as the proportion of loan which a lender bought. We do not have a loan which was overfunded.

```
sum(loans[["funding_rate"]] == 100)
```

```
## [1] 1122521
```

There are 1,122,521 loans which are fully funded at 100%. The rest,

```
sum(loans[["funding_rate"]] < 100)
```

```
## [1] 7217
```

are underfunded. Could these be loans in fundraising stage?

```
length(subset(loans, loans$funding_rate < 100)[['status']])
```

```
## [1] 7217
```

. . . and Yes, they are loans in "fundraising" stage.

One of the objectives of this project was to assess the probability of successful funding for a loan application. The variable *funding_rate* was meant to be our response variable. The dataset contains high success rate with underfunded loans being the only one not fully funded. This makes our objective impossible to achieve.

2.4.8 funding_duration

This variable defines the period (in days) it takes for a loan to be bought by the lenders.

```
loans$funding_duration <- as.Date(loans$planned_expiration_date) - as.Date(loans$posted_date)
loans$funding_duration <- as.integer(loans$funding_duration)
```

```
summary(loans$funding_duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     1.0    30.0    30.0    32.5    30.0  1674.0  368810
```

The NAs are due to some loan applications missing a *planned_expiration_date*. We could set these to have the median value.

```
mean_duration <- mean(loans$funding_duration, na.rm = TRUE)
loans$funding_duration[is.na(loans$funding_duration)] <- mean_duration
summary(loans$funding_duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   30.00   30.00   32.49   32.49 1674.00
```

The max value points to a loan which has a very long duration. Ideally such a high duration is not practical since a loan should not stay for that long at "fundraising stage". If we consider a maximum period of 90 days as the funding period, then 339 observations are affected. 90 days is reasonable because it is a common credit period in the finance world.

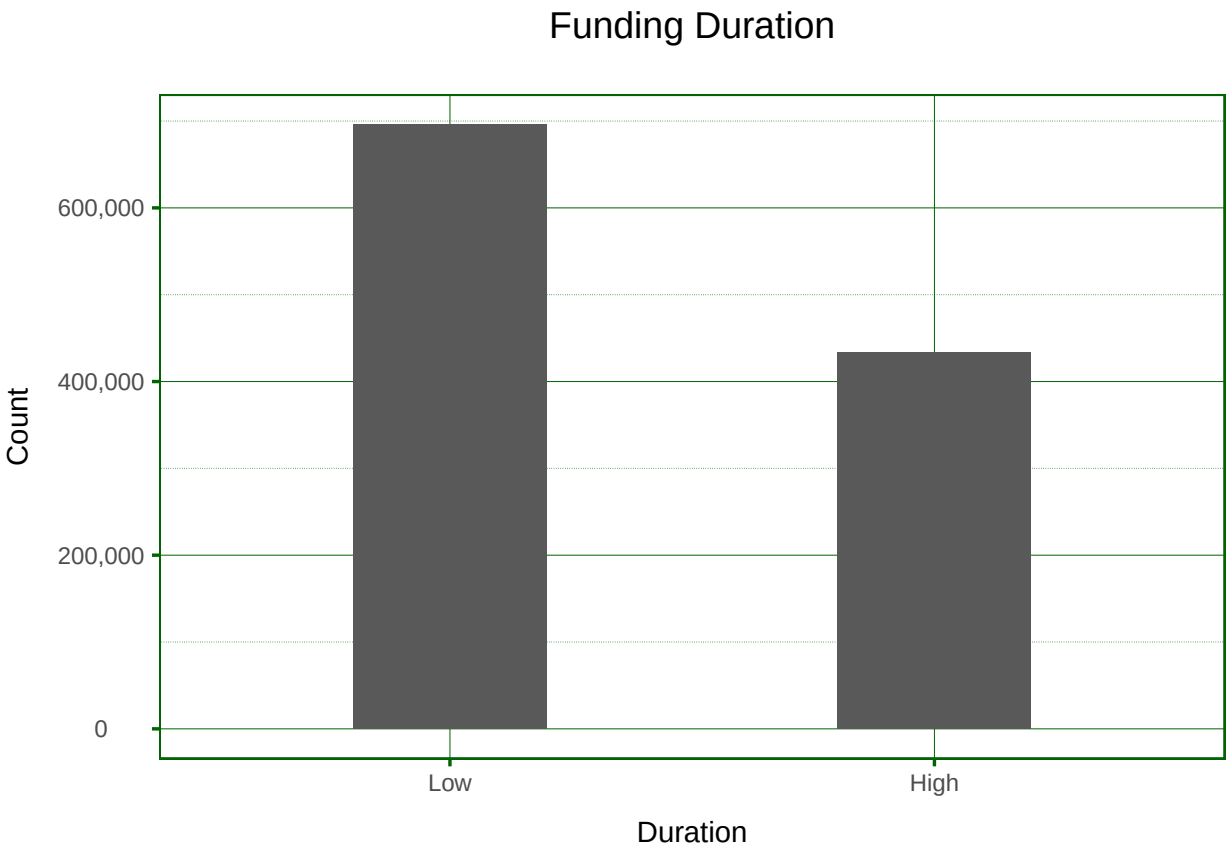
```
sum(loans[["funding_duration"]] > 90)
```

```
## [1] 337
```

Further, since there is very little variablity (median and mean values are very close), we can create groups for the funding duration treating them as categories:

- low = less than or equal to 30 days
- high = more than 30 days

```
loans$funding_duration <- cut(loans$funding_duration, c(0, 30, 1674),
                              labels=c('Low','High'),
                              ordered_result = TRUE,
                              include.lowest = TRUE)
```



More loans are funded within 30 days. Perhaps this goes with the thought that most loan amounts are small thereby they do not need a lot of time to seek funding.

2.4.9 country

Loan applicants register their country when they apply for loans.

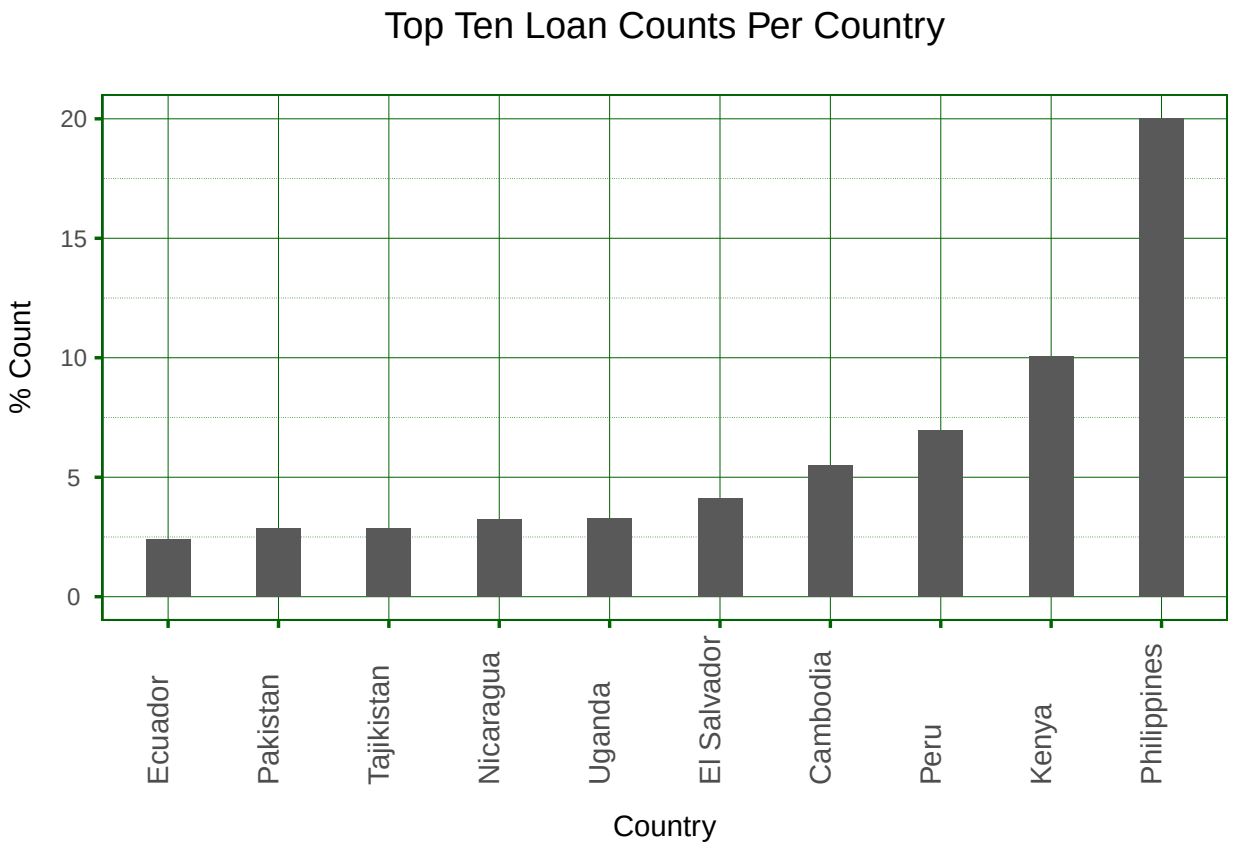
```
loans$country <- factor(loans$location.country)
# table(loans$country)
```

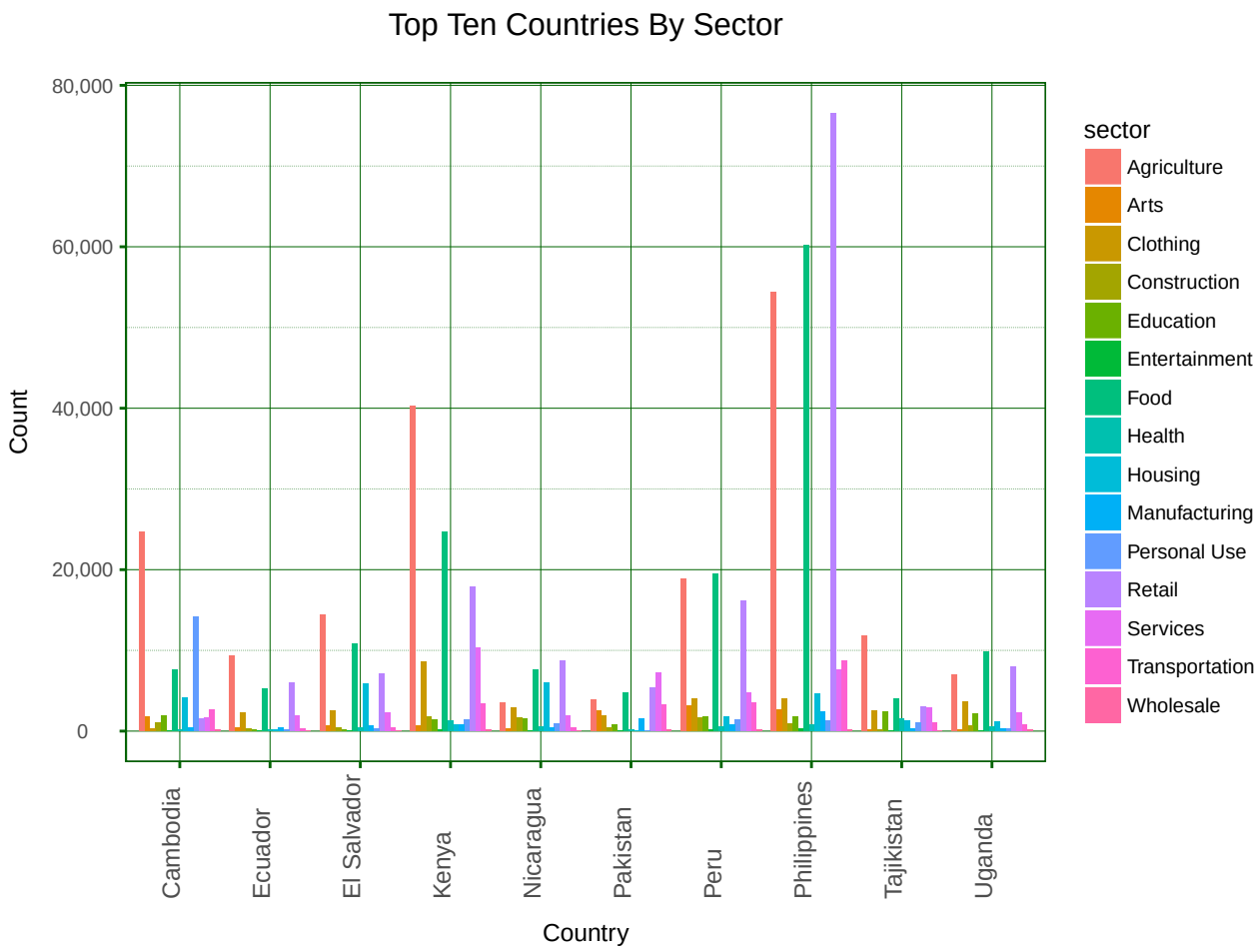
```
country_dist <- data.frame(sort(table(loans$country)))
country_dist$Percentage <- country_dist$Freq / sum(country_dist$Freq) * 100
colnames(country_dist) <- c('Country', 'Frequency', 'Percentage')
# country_dist
```

Kenya and Phillipines have the high counts for borrowers. Perhaps it would be intersting to complement this dataset with one with economic ranking for countries to understand if there is a relationship.

The top ten countries by loan count: 1. Phillipines (19.4%) 2. Kenya (10.0%) 3. Peru (6.78%) 4. Cambodia (5.34%) 5. El Salvador (4.53%) 6. Uganda (3.34%) 7. Nicaragua (3.28%) 8. Tajikistan (2.98%) 9. Pakistan (2.86%) 10. Equador (2.38%)

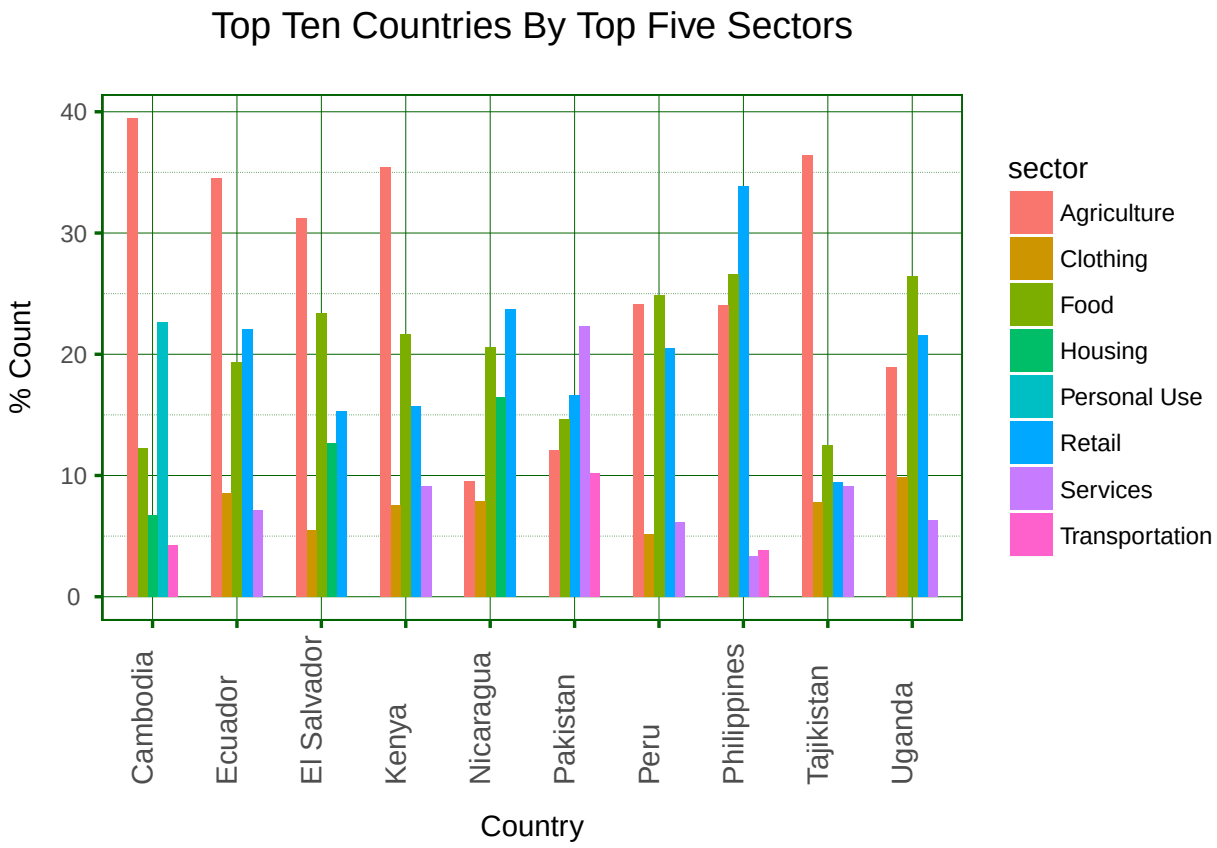
The top ten countries account for 60.89% of the total loans.


Top Ten Loan Counts Per Country
% Count
20
15
10
5
0
Ecuador
Pakistan
Tajikistan
Nicaragua
Uganda
El Salvador
Cambodia
Peru
Kenya
Philippines
Country



Lets plot the top 5 sectors per country.

```
loans_top_ten_grouped <- group_by(loans_top_ten, country, sector)
per_sector <- loans_top_ten_grouped %>%
    summarise(count_per_sector=n()) %>%
    mutate(prop = 100*count_per_sector/sum(count_per_sector)) %>%
    arrange(country, desc(count_per_sector)) %>%
    slice(seq(5))
#per_sector
```



Sectors in the plot above attract most of the microlending spend. They can be viewed as sectors through which poverty reduction efforts are addressed.

Sectors present in all top ten countries are "Agriculture" and "Food". "Retail" closely follows missing only in Cambodia. "Agriculture" tops in 5 countries (Cambodia, Equador, El Savaldor, Kenya and Tajikistan) followed by "Retail" in 2 countries (Phillipines and Nicaragua), "Food" in 2 countries (Peru and Uganda) and "Services" in 1 country (Pakistan).

2.4.10 description languages

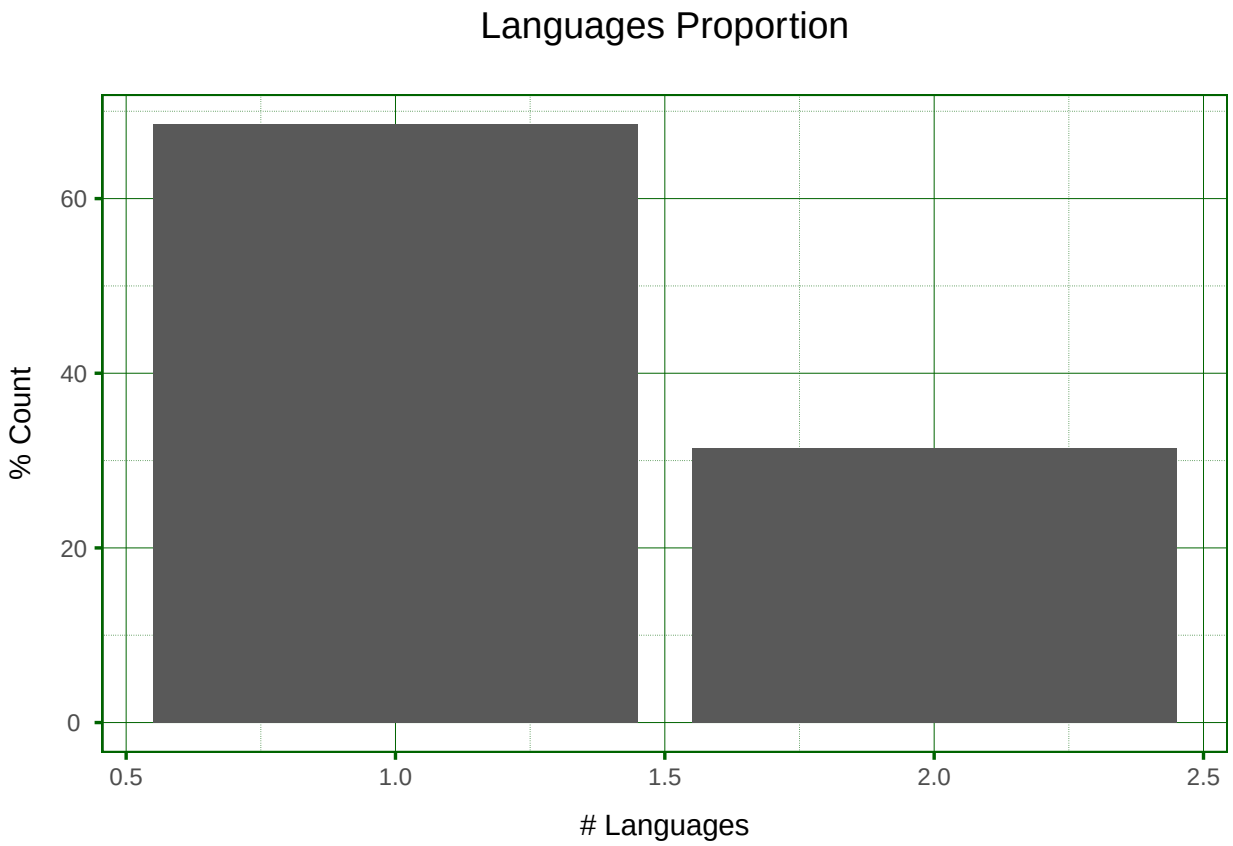
These are languages borrowers choose to use for their loan applications.

```
loans$num_languages <- str_count(loans$description.languages, ',') + 1
```

```
summary(loans$num_languages)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.315   2.000   2.000
```

```
loans_prop <- loans %>%
    group_by(num_languages) %>%
    summarise(count=n()) %>%
    mutate(percent=100*count/sum(count))
```

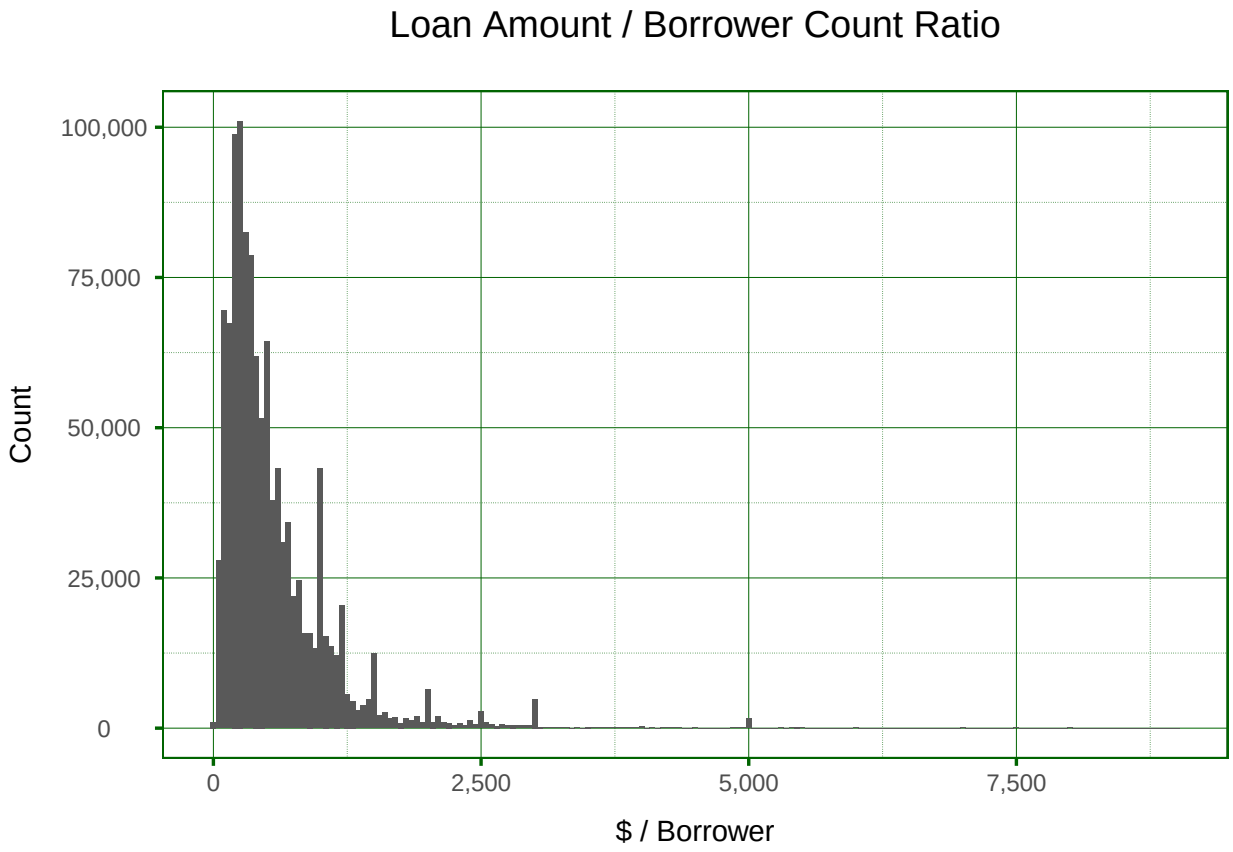


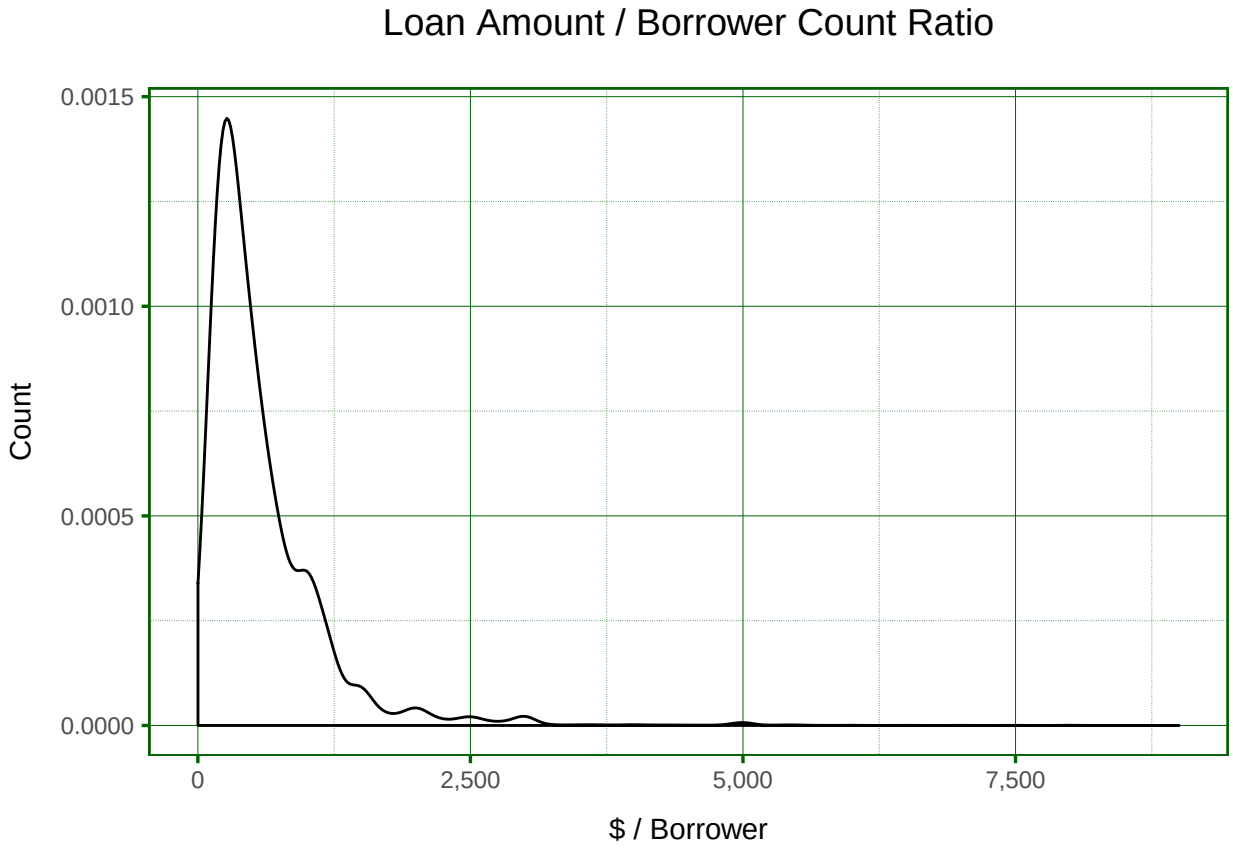
Most borrower profiles have only one language selected. This is about twice as much as the single language profiles - about 68% for single language profile viz 30% for profiles with two languages. It is of interest checking what languages if any the lenders have on their profiles and if they influence decision to buy loans (loans whose profile owners have same language as themselves).

2.4.11 loan_amount vs borrower_count

```
loans$amt_borrower_ratio <- loans$loan_amount / loans$borrower_count
summary(loans$amt_borrower_ratio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.19  250.00  400.00  580.40  725.00 9000.00
```


Loan Amount / Borrower Count Ratio
100,000
75,000
50,000
25,000
0
Count
0
2,500
5,000
7,500
$ / Borrower


Loan Amount / Borrower Count Ratio
0.0015
0.0010
0.0005
0.0000
Count
0
2,500
5,000
7,500
$ / Borrower

Loan amount per borrower assumes a right skewed shape - there is a longtail to the right. Most of the borrowers seek small amount loans and there are also borrowers with high dollar amounts. From the summary calculation, the average amount per borrower is $580 while 75% are less than $725. The other 25% are high loan amounts up to a maximum of $9,000. The smoothened density plot show the unimodal nature of the data with a mode of about $300.

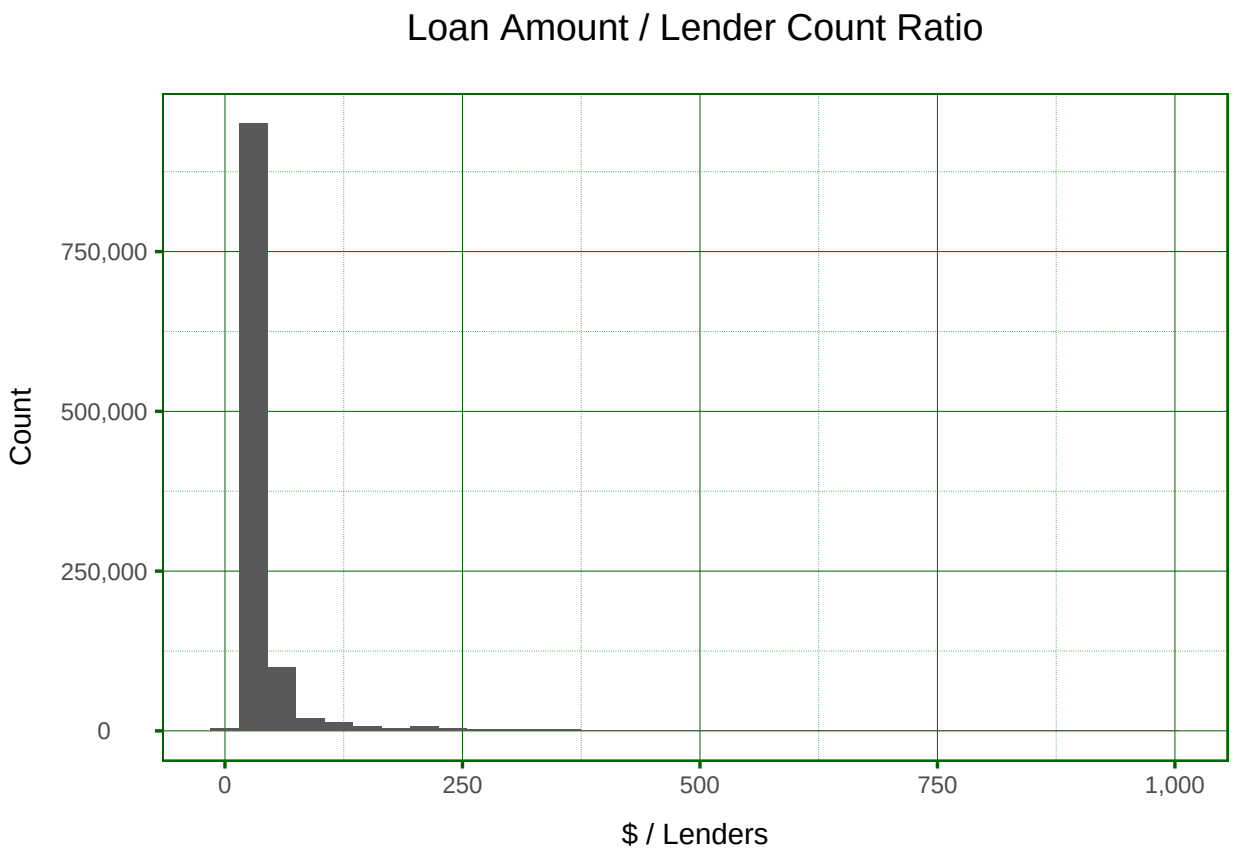
2.4.12 loan_amount vs lender_count

```
loans_with_lenders <- subset(loans, lender_count > 0)
loans_with_lenders$amt_lender_ratio <- loans_with_lenders$loan_amount / loans_with_lenders$lender_count
summary(loans_with_lenders$amt_lender_ratio)
```

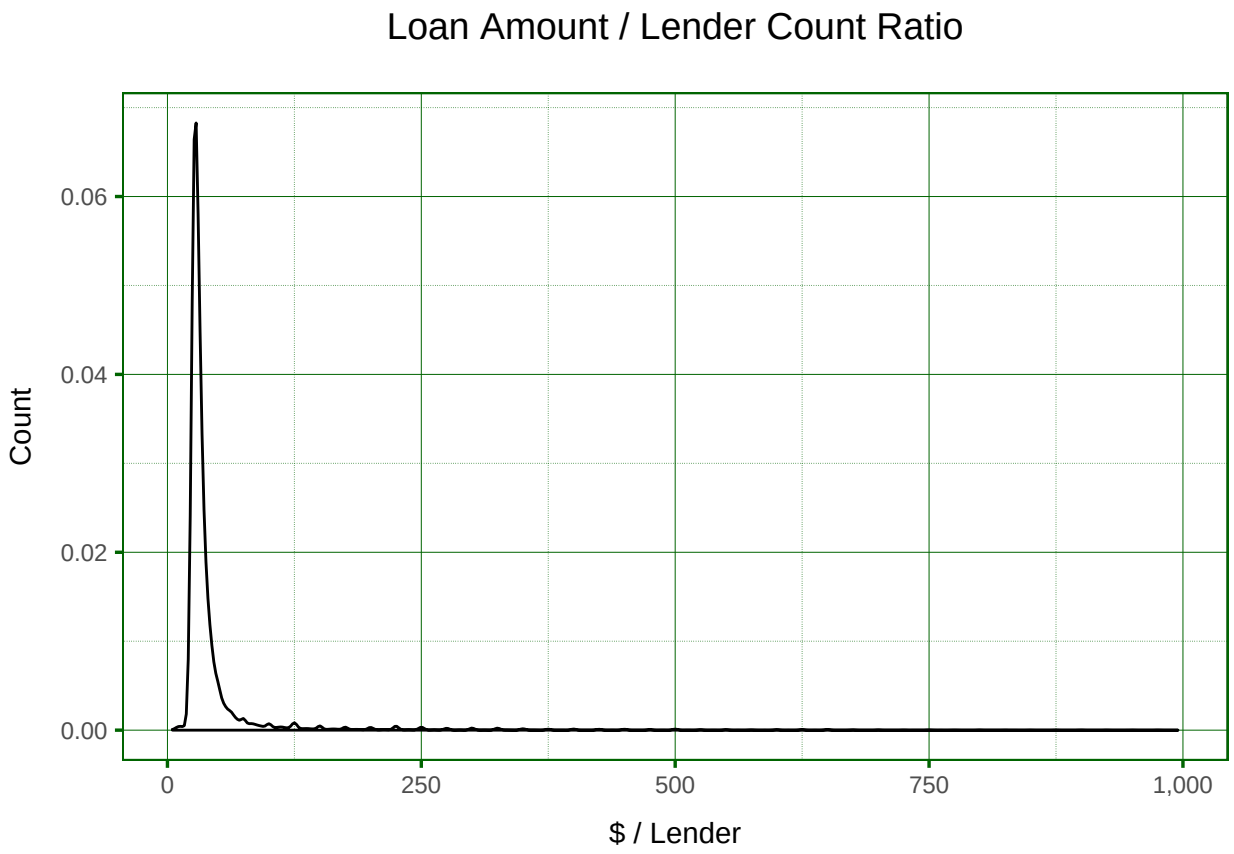
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   26.67   30.00   50.30   37.50 8775.00
```

The maxmimum amount though is quite out of the ordinary, looks like some lenders made huge investments. We adjust for this.

```
loans_with_lenders <- subset(loans_with_lenders, amt_lender_ratio < 1000)
```




```
ggplot(loans_with_lenders, aes(x=amt_lender_ratio)) +
    geom_density(adjust = 5) +
    scale_x_continuous(labels = comma) +
    scale_y_continuous(labels = comma) +
    labs(x = "$ / Lender", y = "Count", title = "Loan Amount / Lender Count Ratio")
```



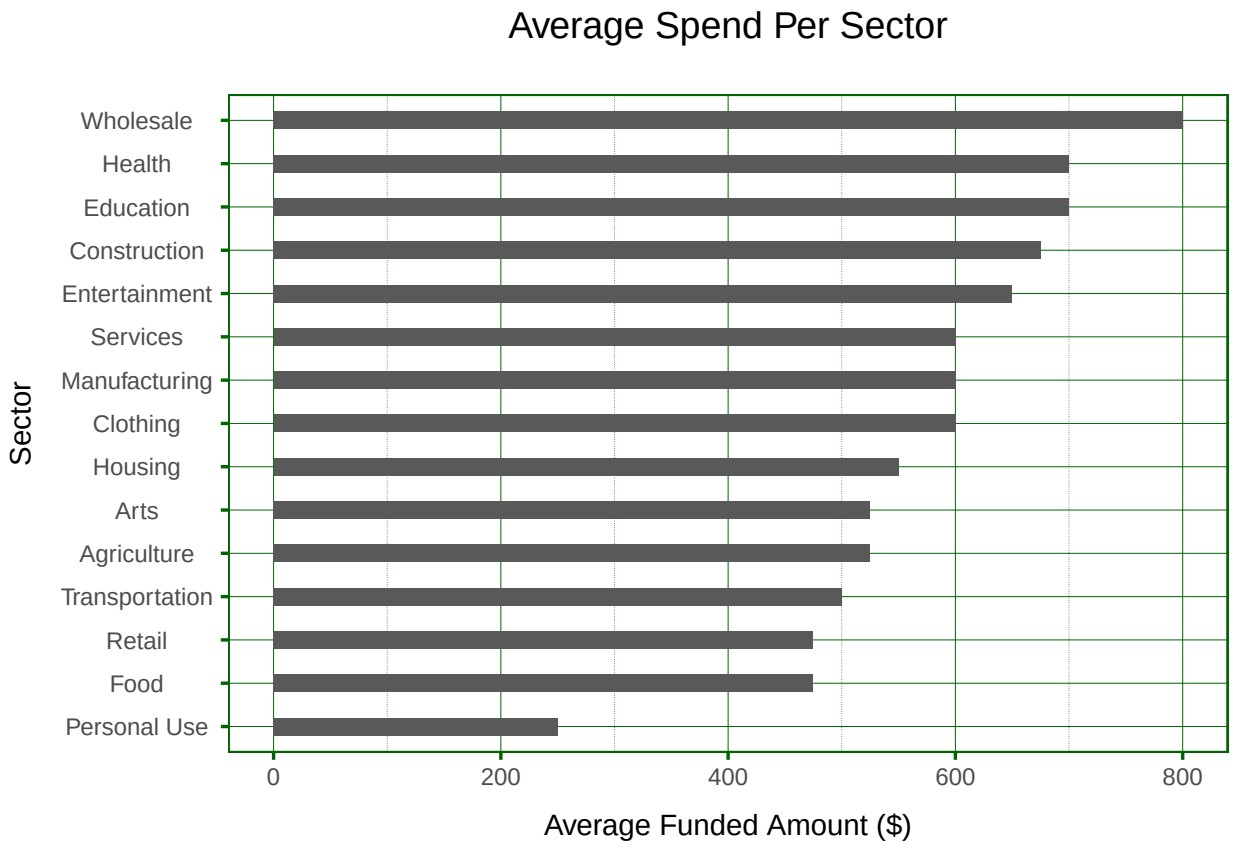
THe average Dollar amount per lender is $30. This means many more lenders combine effort to satisfy a loan bid. Previously we saw an average loan amount of $580 leading us to conclude that on average a loan requires at least 19 lenders and points to the fact that lenders tend to spread their risk by buying multiple loans and putting in small amounts.

The lender ratio has a very narrow range beyond which it assumes huge values. Lenders tend to either place small amounts when they buy loans while others place big amounts. Majority of them place small amounts.

3. Other Analysis

Loan awarded per sector

```
median_per_sector <- group_by(loans, sector) %>%
    summarise(med=median(funded_amount))
#median_per_sector <- arrange(median_per_sector, desc(med))
```



"Wholesale" sector has the highest median value. Together with "Construction", "Education", "Entertainment" and "Health", their median values equal or exceed $600. Typically, these sectors demand high dollar amounts - like doing a housing project or funding medication can be very expensive.

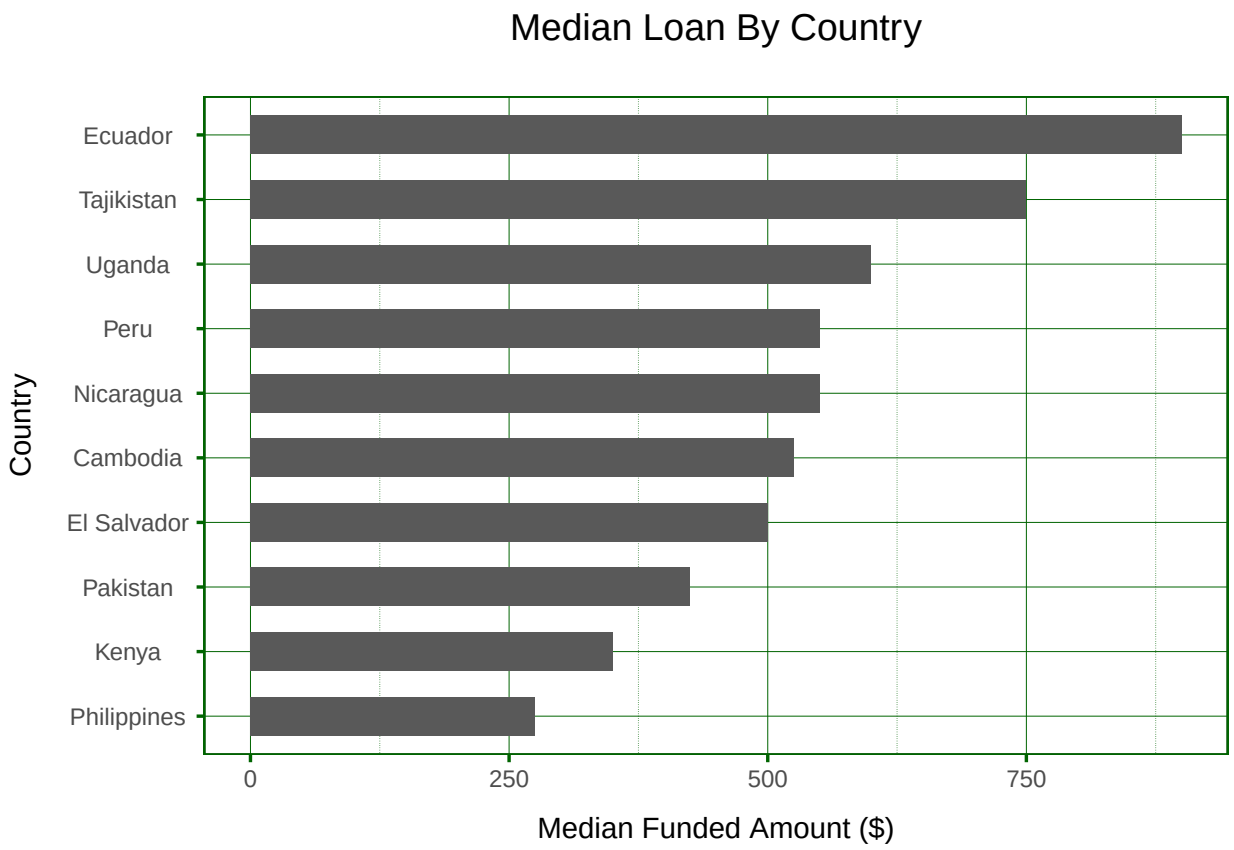
On the other hand, highly frequent sectors like "Agriculture", "Food" and "Retail" attract small investment amounts - like small retail groceries business and small scale farming.

Loan Distribution By Country

```
median_per_country <- group_by(loans, country) %>% summarise(med=median(funded_amount))
```

Botswana stands out as having the highest median value while it has one loan recorded. This probably means a better way to look at this is to narrow down to the previously identified top ten countries.

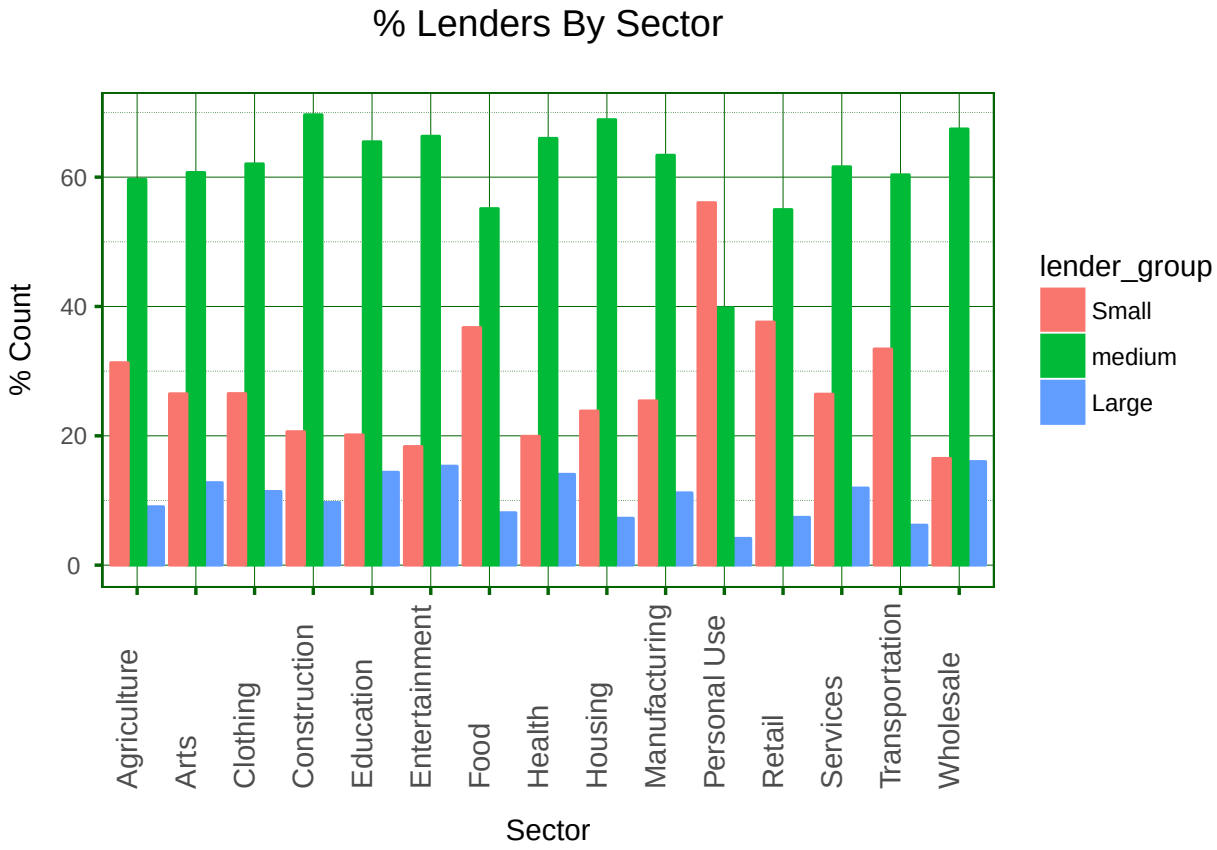
```
top_ten <- loans %>% filter(country %in% country_top_ten$Country)
median_per_country <- group_by(top_ten, country) %>% summarise(med=median(funded_amount))
```



Previously we saw that "Agriculture", "Food" and "Retail" are sectors that dominate the top ten countries.

Lender Distribution Per Sector

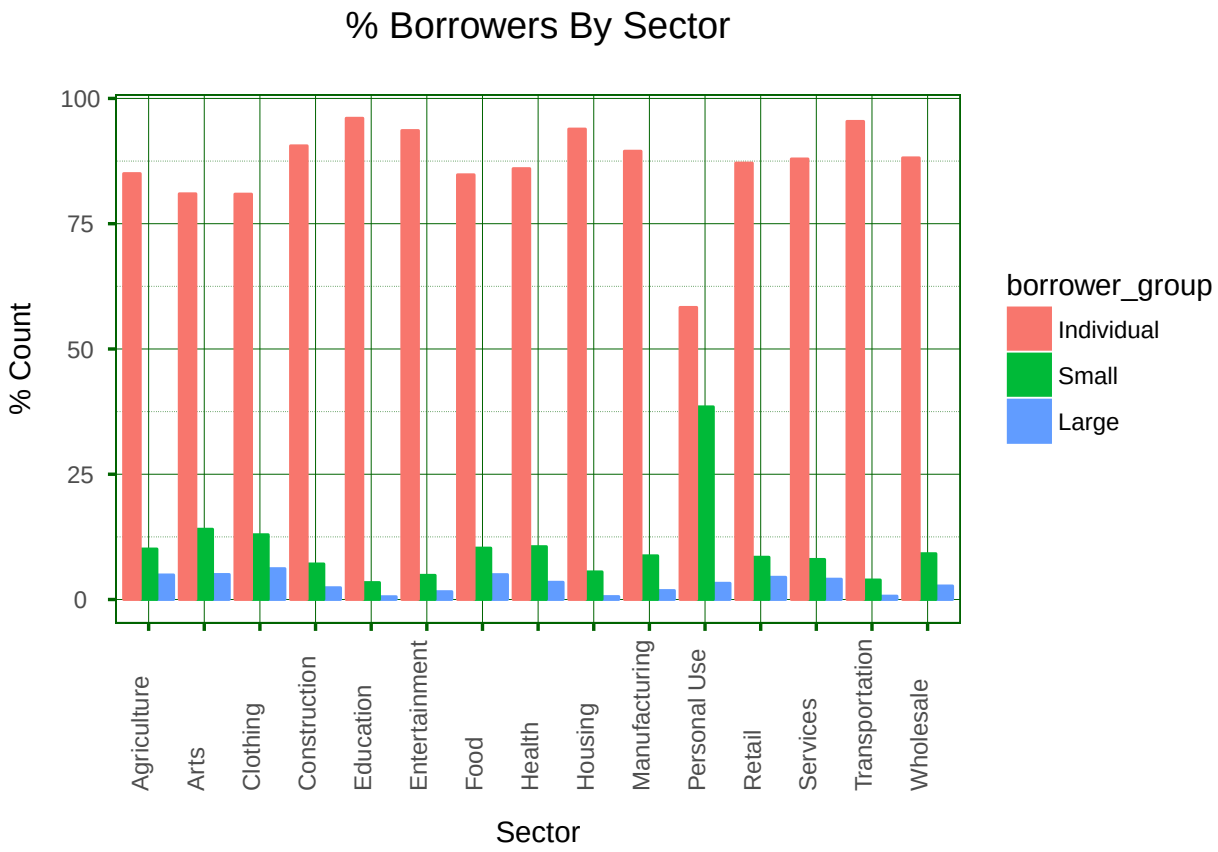
```
lender_prop <- loans %>%
    group_by(sector, lender_group) %>%
    summarise(n=n()) %>%
    mutate(percent=100*n/sum(n))
```



"medium" group of lenders dominates across all sectors implying that almost all loans across all sectors have 10 - 50 lenders. This adds weight to our postulation that a loan requires 19 lenders to buy it.

For borrowers,

```
borrower_prop <- loans %>%
    group_by(sector, borrower_group) %>%
    summarise(n=n()) %>%
    mutate(percent=100*n/sum(n))
```



Individual borrowers are dominate all sectors, even in the least favorite sectors like "Entertainment". Large group of borrowers are less in all the sectors. People like to run projects on their own - sole proprietorship model of business is evident here and it is a common trait of microenterprises.

Time Aspect of the Loans

Let us now evaluate loans trend over time. We need to compute by year, month and day.

NOTE: The data set covers only upto mid December 2016 when the data was fetched.

```
max(loans$posted_date)
```

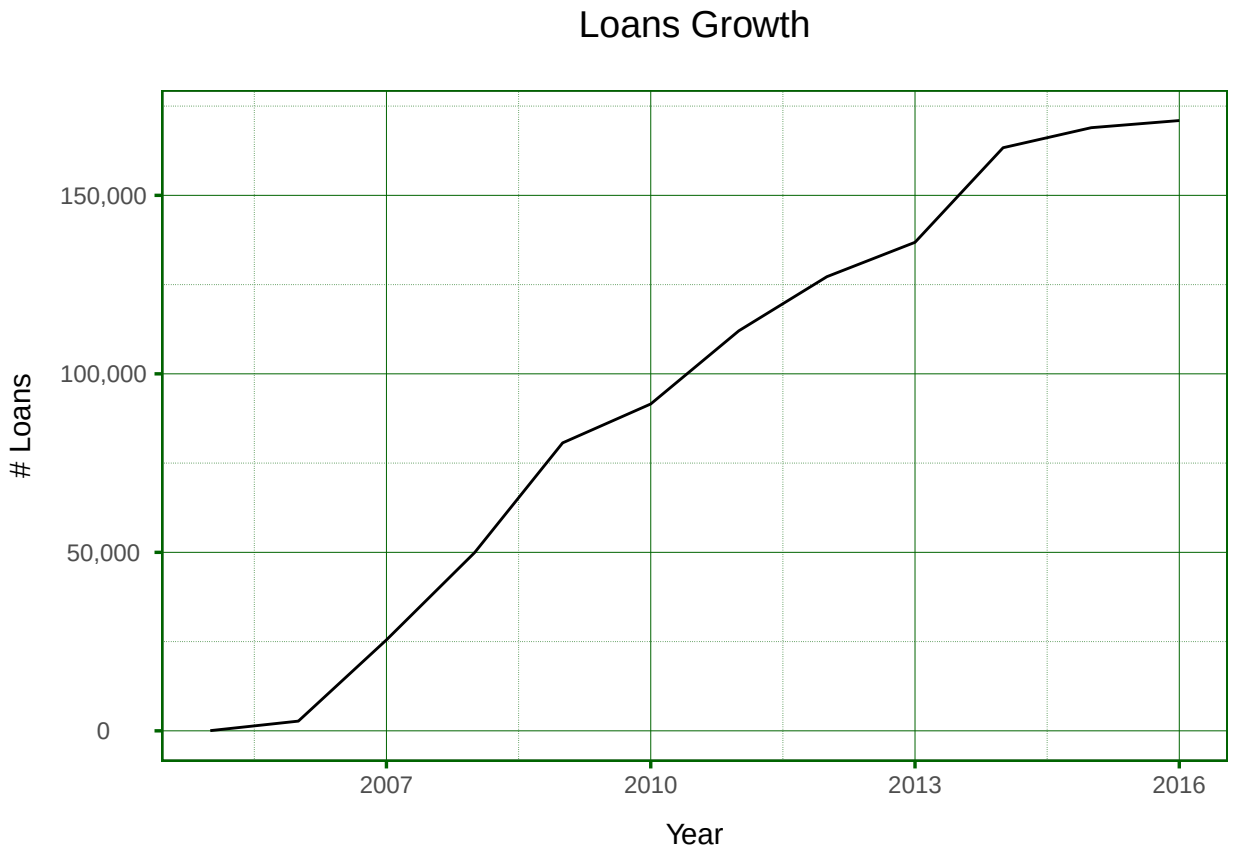
```
## [1] "2016-12-13"
```

```
loans <- loans %>% mutate(year=as.numeric(format(posted_date, format = "%Y")),
                          month=as.numeric(format(posted_date, format = "%m")),
                          day=as.numeric(format(posted_date, format = "%d")))
```

```
trend_year <- loans %>% group_by(year) %>% summarise(n=n())
trend_year
```

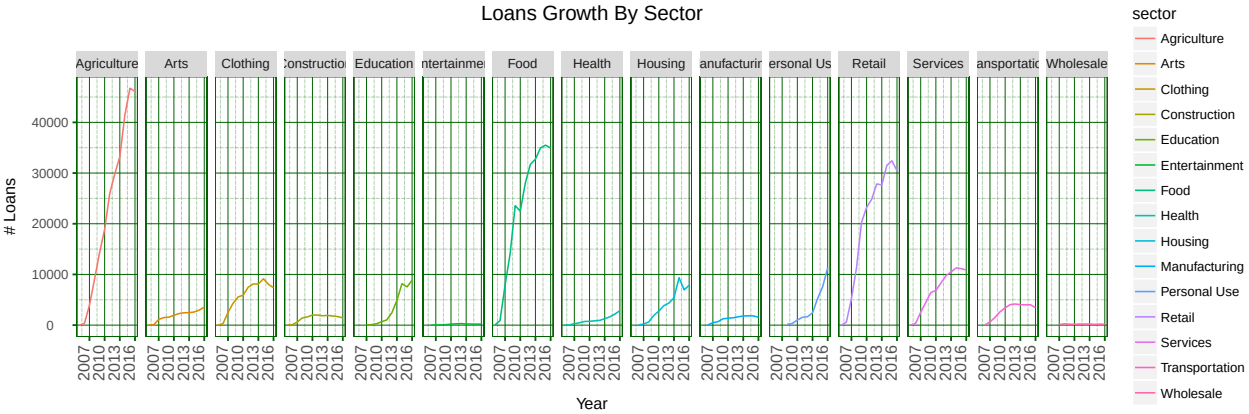
```
## # A tibble: 12 × 2
##     year      n
##    <dbl>  <int>
## 1   2005     55
## 2   2006   2722
## 3   2007  25472
## 4   2008  49910
## 5   2009  80660
```

```
## 6   2010  91562
## 7   2011 112053
## 8   2012 127214
## 9   2013 136847
## 10  2014 163340
## 11  2015 168951
## 12  2016 170952
```



Kiva loans facility has grown over time since 2005, the earliest year in the data. It administered 55 loans in the year 2005 and did 170k in the last year. The rate of growth is however descreasing since 2014. Could this be due to competition from other microlenders?

```
trend_year_sector <- loans %>% group_by(year, sector) %>% summarise(n=n())
#trend_year_sector
```



"Agriculture", "Food" and "Retail" have rapidly grown over the years. "Personal Use" was introduced last and has a pretty rapid growth. "Housing", "Education" and "Food" had brief negative growth at different times perahps due to fluctuation in building materials prices, access to state student loans or weather conditions respectively. "Wholesale" and "Entertainment" sectors have flat growth.

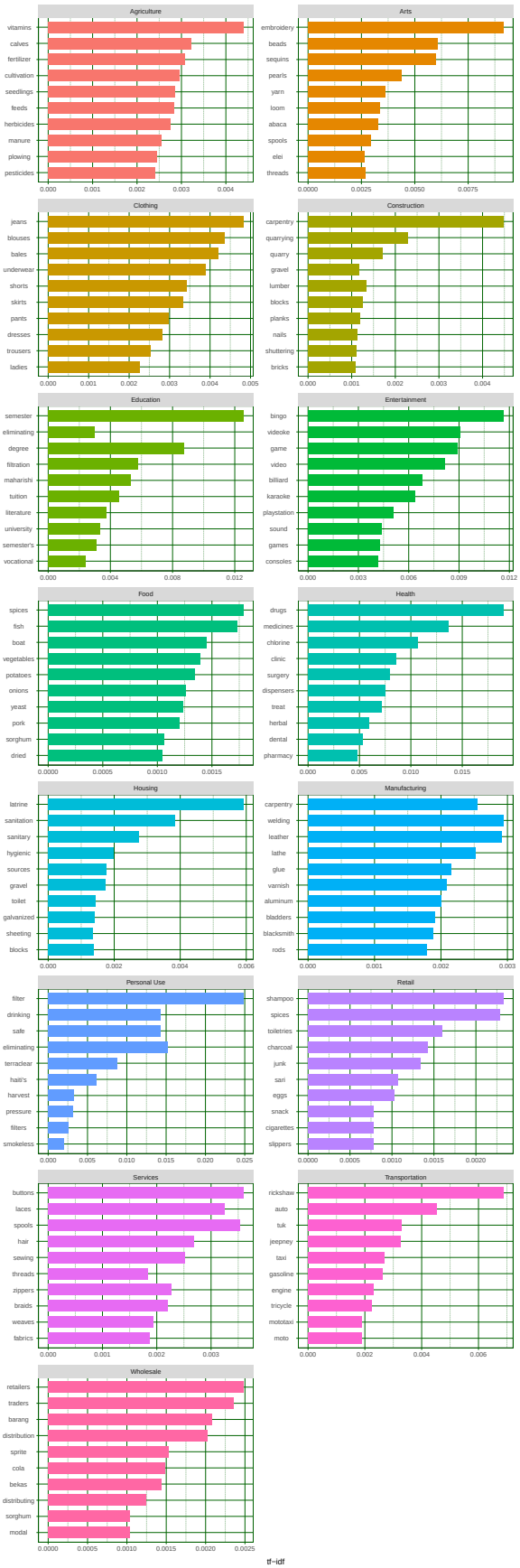
Analyzing each sector's descriptions:

```
loan_use <- subset(loans, select=c("sector", "use"))
data('stop_words')
my_stop_words <- data_frame(word=c("buy", "purchase", "i.e", "sell"))
word_count <- loan_use %>%
    unnest_tokens(word, use, drop=FALSE) %>%
    anti_join(stop_words) %>%
    anti_join(my_stop_words) %>%
    count(sector, word) %>%
    bind_tf_idf(word, sector, n) %>%
    group_by(sector) %>%
    top_n(10, tf_idf) %>%
    mutate(word = reorder(word, tf_idf))
```

```
## Joining, by = "word"
## Joining, by = "word"
```

```
head(word_count)
```

```
## Source: local data frame [6 x 6]
## Groups: sector [1]
##
##        sector        word     n          tf        idf      tf_idf
##        <fctr>      <fctr> <int>       <dbl>      <dbl>       <dbl>
## 1 Agriculture      calves  4517 0.004221929 0.76214005 0.003217701
## 2 Agriculture cultivation  5050 0.004720111 0.62860866 0.002967103
## 3 Agriculture       feeds  9767 0.009128976 0.31015493 0.002831397
## 4 Agriculture  fertilizer 47623 0.044512053 0.06899287 0.003071014
## 5 Agriculture  herbicides  3222 0.003011525 0.91629073 0.002759432
## 6 Agriculture      manure  8777 0.008203647 0.31015493 0.002544402
```


Agriculture
vitamins
calves
fertilizer
cultivation
seedlings
feeds
herbicides
manure
plowing
pesticides
0.000
0.001
0.002
0.003
0.004
Arts
embroidery
beads
sequins
pearls
yarn
loom
abaca
spools
elei
threads
0.0000
0.0025
0.0050
0.0075
Clothing
jeans
blouses
bales
underwear
shorts
skirts
pants
dresses
trousers
ladies
0.000
0.001
0.002
0.003
0.004
0.005
Construction
carpentry
quarrying
quarry
gravel
lumber
blocks
planks
nails
shuttering
bricks
0.000
0.001
0.002
0.003
0.004
Education
semester
eliminating
degree
filtration
maharishi
tuition
literature
university
semester's
vocational
0.000
0.004
0.008
0.012
Entertainment
bingo
videoke
game
video
billiard
karaoke
playstation
sound
games
consoles
0.000
0.003
0.006
0.009
0.012
Food
spices
fish
boat
vegetables
potatoes
onions
yeast
pork
sorghum
dried
0.0000
0.0005
0.0010
0.0015
Health
drugs
medicines
chlorine
clinic
surgery
dispensers
treat
herbal
dental
pharmacy
0.000
0.005
0.010
0.015
Housing
latrine
sanitation
sanitary
hygienic
sources
gravel
toilet
galvanized
sheeting
blocks
0.000
0.002
0.004
0.006
Manufacturing
carpentry
welding
leather
lathe
glue
varnish
aluminum
bladders
blacksmith
rods
0.000
0.001
0.002
0.003
Personal Use
filter
drinking
safe
eliminating
terraclear
haiti's
harvest
pressure
filters
smokeless
0.000
0.005
0.010
0.015
0.020
0.025
Retail
shampoo
spices
toiletries
charcoal
junk
sari
eggs
snack
cigarettes
slippers
0.0000
0.0005
0.0010
0.0015
0.0020
Services
buttons
laces
spools
hair
sewing
threads
zippers
braids
weaves
fabrics
0.000
0.001
0.002
0.003
Transportation
rickshaw
auto
tuk
jeepney
taxi
gasoline
engine
tricycle
mototaxi
moto
0.000
0.002
0.004
0.006
Wholesale
retailers
traders
barang
distribution
sprite
cola
bekas
distributing
sorghum
modal
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
tf−idf

From top keywords for "Agriculture" we infer that them main agricultural activities are cattle rearing and planting crop. For "Food" we can generally categorize items under fresh produce (vegetables, potatoes, onion), grains (sorghum, yeast), meat products (pork, fish) and spices. "Retail" on the other hand has items like shampoo, toiletries, slippers, spices, eggs, snack, charcoal and cigarette. From the analysis of the data, Kenya and Phillipines are the top countries. I can attest the keywords here are fairly good representation of what happens in my country.

We also previously observed a recent rapid growth in the "Personal Use" sector. The keywords in this sector are filter, drinking and safe. These keywords are tied to safe drinking water. It seems like more loan applicants in this sector are looking for means to get clean drinking water. Does it mean water related diseases have been a menace and there is a campaign against them? Most likely this is the case. Chlorine is one of the keywords in the "Health" sector and it is mainly used for water purifiation. On the same breath, in "Housing" sector, latrine and sanitation are the top keywords. Proper sanitation and clean drinking water are main themes in the low income population.

"Education" sector has keywords implying the loans in this sector are mainly used to pay for college tuition in the universities and vocational training centres.

We can also infer that timber, stone and brick are common building materials used for construction.

4. Conclusion

1. Kiva plays a big role in bridging the financing gap for the underprivileged borrowers.
2. From the analysis of the sectors we see how the loans help improve their lives through broad categories under health and sanitation, proper housing and thriving small businesses.
3. The funding rate which stands at over 95% is quite impressive for this group of borrowers who do not use any collateral and entirely depend on the benevolence of the lenders. One of the objectives of this study had been to attempt to predict success in funding for loans. Very few loans are underfunded and this rendered the identified objective unintuitive.

5. Further Work

We can enhance this dataset by combining it with the lenders dataset and trying to uncover further insights in the context of lenders. Also other nice visualizations to add in the analysis would be geo-based visualization and a plot to show flow of money from lender countries to borrower countries and to the sectors.