Midterm Project - Kiva.org

Jordan Lardieri - Boston University

December 09, 2018

##

Abstract

My interest is in the social sciences, and the use of data science for social good. From improving global health to city infrastructure, we can use data to help solve major societal issues. In my career, I'd like to contribute to these efforts.

I chose my project because it aligns with this goal. I'm interested in *Kiva.org's* cause because they empower so many individuals and communities around the world through crowdsourced loans, not donations. Anyone can go to the website, get familiar with someone's story and make a loan contribution.

Their website has a number of focus areas; women, single parents, conflict zones, water, and education to name a few. As a nonprofit, Kiva's mission is to connect people through lending to help alleviate poverty. Kiva supports 3 million borrowers in more than 80 countries, creating opportunities for individuals, their families, and their communities.

My project aims to take a deeper look at the Kiva loan data, to see if there are any underlying themes and behaviors that differ between regions and countries. Specifically, is there variation in category and amount of funding and to whom.

Figure 1: Kiva.org.

Method

The dataset is sourced and merged from 2 Kaggle datasets; loan detail from *Kiva.org* and regional multidimenisonal poverty index (MPI) detail. It was reshaped to have tag information per row (tags are additional information provided for borrowers, for example, "Elderly","Woman Owned Biz").

Variables *Partner.ID, Country, Region, WorldRegion, Sector, Activity, Use, Tags, Gender, Tags, Date, FundedLoan A*

Note that there is more data to be merged that will be relevant to analysis, loan theme by region, human development index, and population below the poverty line in the future.

Dataset 1: Kiva Loans

The dataset contains the majority of the loan detail provided by Kiva.

```
##    partner_id funded_amount loan_amount           sector
## 1         247          2225        2225           Retail
## 2         334           250         250         Services
## 3         334           200         200      Agriculture
## 4         334           150         150   Transportation
## 5         334           250         250     Construction
## 6         334           250         250      Agriculture
##                       activity  country     region
## 1 Personal Products Sales Pakistan      Lahore
## 2                    Sewing     India Maynaguri
## 3                     Dairy     India Maynaguri
## 4             Transportation     India Maynaguri
## 5    Construction Supplies     India Maynaguri
## 6                   Farming     India  Dhupguri
##                                              borrower_genders
## 1 female,female,female,female,female,female,female,female
## 2                                                   female
## 3                                                   female
## 4                                                   female
## 5                                                   female
## 6                                                   female
##                                                                         use
## 1                                          to buy hair oils to sell.
## 2                                    to purchase a sewing machine.
## 3              To purchase a dairy cow and start a milk products business .
## 4  To repair their old cycle-van and buy another one to rent out as a source of income
## 5 to purchase stones for starting a business supplying stones to building contractors.
## 6                                    to cultivate broad beans on her own land.
##   term_in_months        date month year
## 1             11 2014-01-01    01 2014
## 2             43 2014-01-01    01 2014
```

```
## 3                 43 2014-01-01     01 2014
## 4                 43 2014-01-01     01 2014
## 5                 43 2014-01-01     01 2014
## 6                 43 2014-01-01     01 2014
##                                       tags Female.count Male.count
## 1 #Parent, #Repeat Borrower, user_favorite            8          0
## 2              user_favorite, user_favorite           1          0
## 3              user_favorite, user_favorite           1          0
## 4              user_favorite, user_favorite           1          0
## 5              user_favorite, user_favorite           1          0
## 6              user_favorite, user_favorite           1          0
##   Woman.Biz M.F.loan F.loan M.loan F.only.loan M.only.loan fully.funded
## 1         0        0      1      0           1           0            1
## 2         0        0      1      0           1           0            1
## 3         0        0      1      0           1           0            1
## 4         0        0      1      0           1           0            1
## 5         0        0      1      0           1           0            1
## 6         0        0      1      0           1           0            1
##   unfunded partial.funded Gender Gender.Var Water
## 1        0              0 female          1     0
## 2        0              0 female          1     0
## 3        0              0 female          1     0
## 4        0              0 female          1     0
## 5        0              0 female          1     0
## 6        0              0 female          1     0
```

Dataset 2: Multidimensional Poverty Index (MPI) and World Region Detail

This dataset contains World Region and MPI variables.

```
##        country    region world_region   MPI
## 1 Afghanistan Badakhshan   South Asia 0.387
## 2 Afghanistan    Badghis   South Asia 0.466
## 3 Afghanistan    Baghlan   South Asia 0.300
## 4 Afghanistan      Balkh   South Asia 0.301
## 5 Afghanistan     Bamyan   South Asia 0.325
## 6 Afghanistan   Daykundi   South Asia 0.313
```

Datasets Merged

This is one iteration of the dataset. I summarized loans and MPI separately by world region, region, and country for analysis.

```
## # A tibble: 6 x 4
## # Groups:   country [6]
```

```
##   country      MPI_country sumloan_amount sumfunded_amount
##   <chr>              <dbl>          <dbl>            <dbl>
## 1 Afghanistan        0.310          0.014            0.014
## 2 Burundi            0.412          2.28             2.17
## 3 Benin              0.320          0.05             0.05
## 4 Burkina Faso       0.548          2.7              2.64
## 5 Belize             0.0201         0.078            0.078
## 6 Brazil             0.0273         0.596            0.595
```
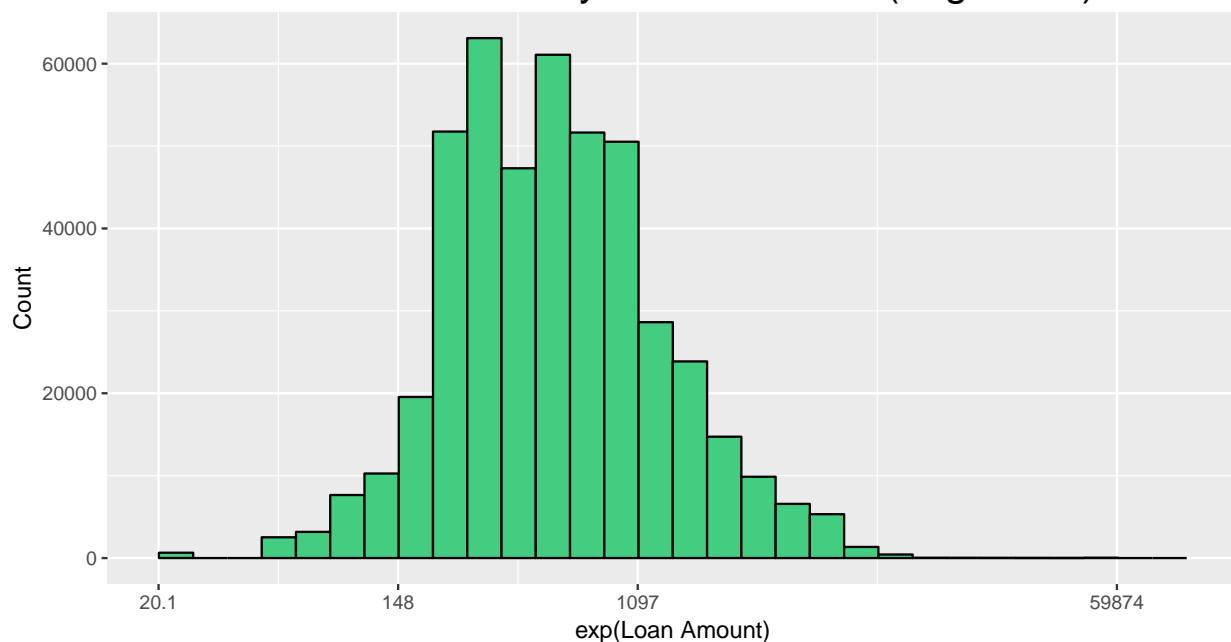
The Loans: Distribution of Fully Funded, Partially Funded, and Unfunded Loans

Not all loans receive full funding.

```
## [1] "No. of Fully Funded Loans = 423089"

## [1] "No. of Unfunded Loans = 2054"

## [1] "No. of Partially Funded Loans = 37022"

## [1] "No. of Over Funded Loans = 2"
```



Distribution of Fully Funded Loans (Log Scale)

In the dataset, an unfunded loan is $FundedAmount$=$0, a partially funded loan is funded amount is less than loan amount and fully funded loan is funded amount=loan amount. Over 91% of Kiva loans are fully funded! From the histogram, the fully funded amounts in the dataset range from about 20 USD to about 5,9875 USD with an average of about 1,100 USD. How does this breakdown by *Country* and *Region*? Is funding going to the most impoverished areas? What about *Gender*? And *Sector*?

5

Frequently Funded Countries and Regions

Top 5 Funded Countries per Year
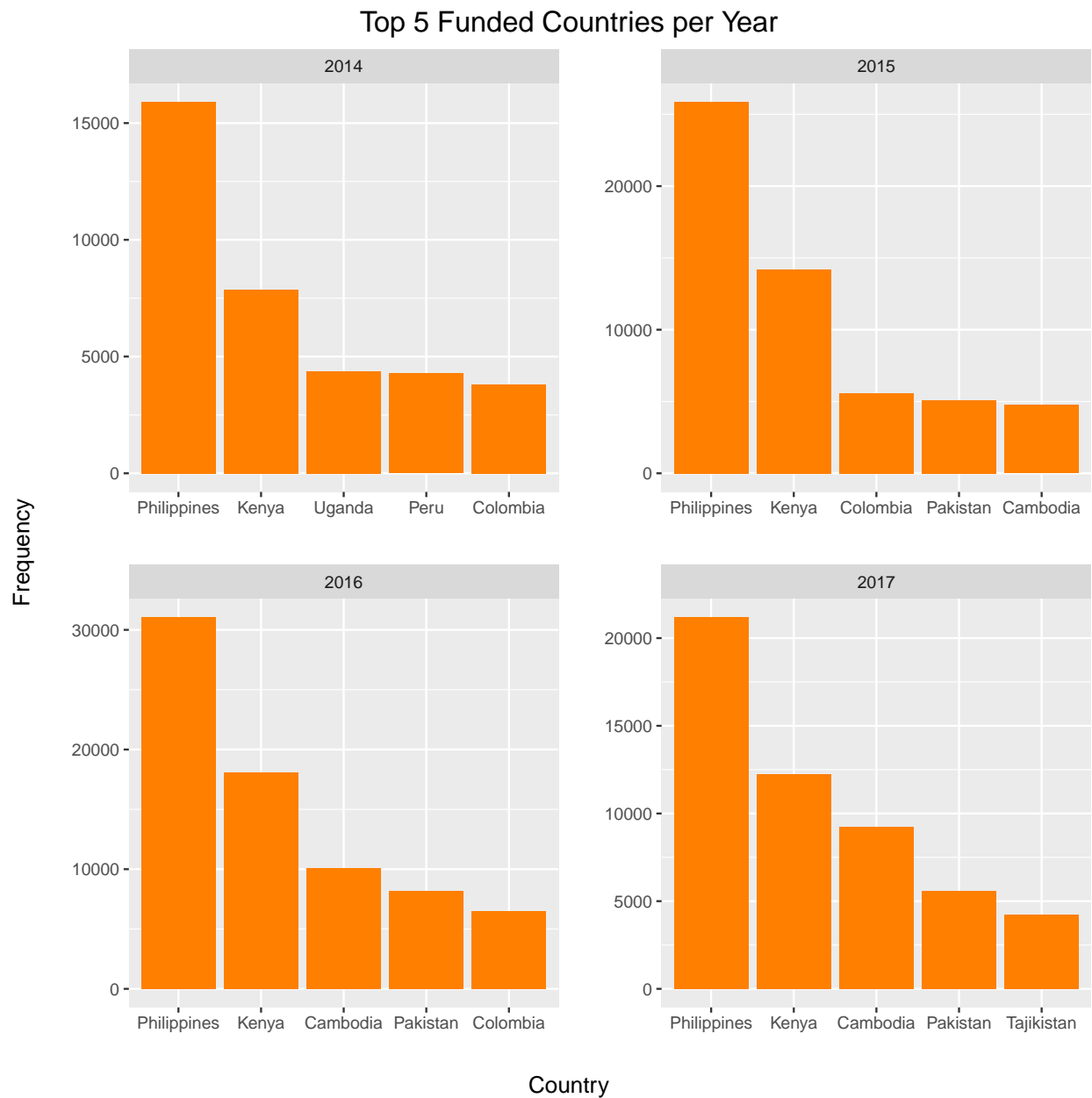
From the bar plots, the top funded countries are consistently the Philippines and Kenya each year, Cambodia is also frequntly funded.

Top 5 Funded Regions per Year
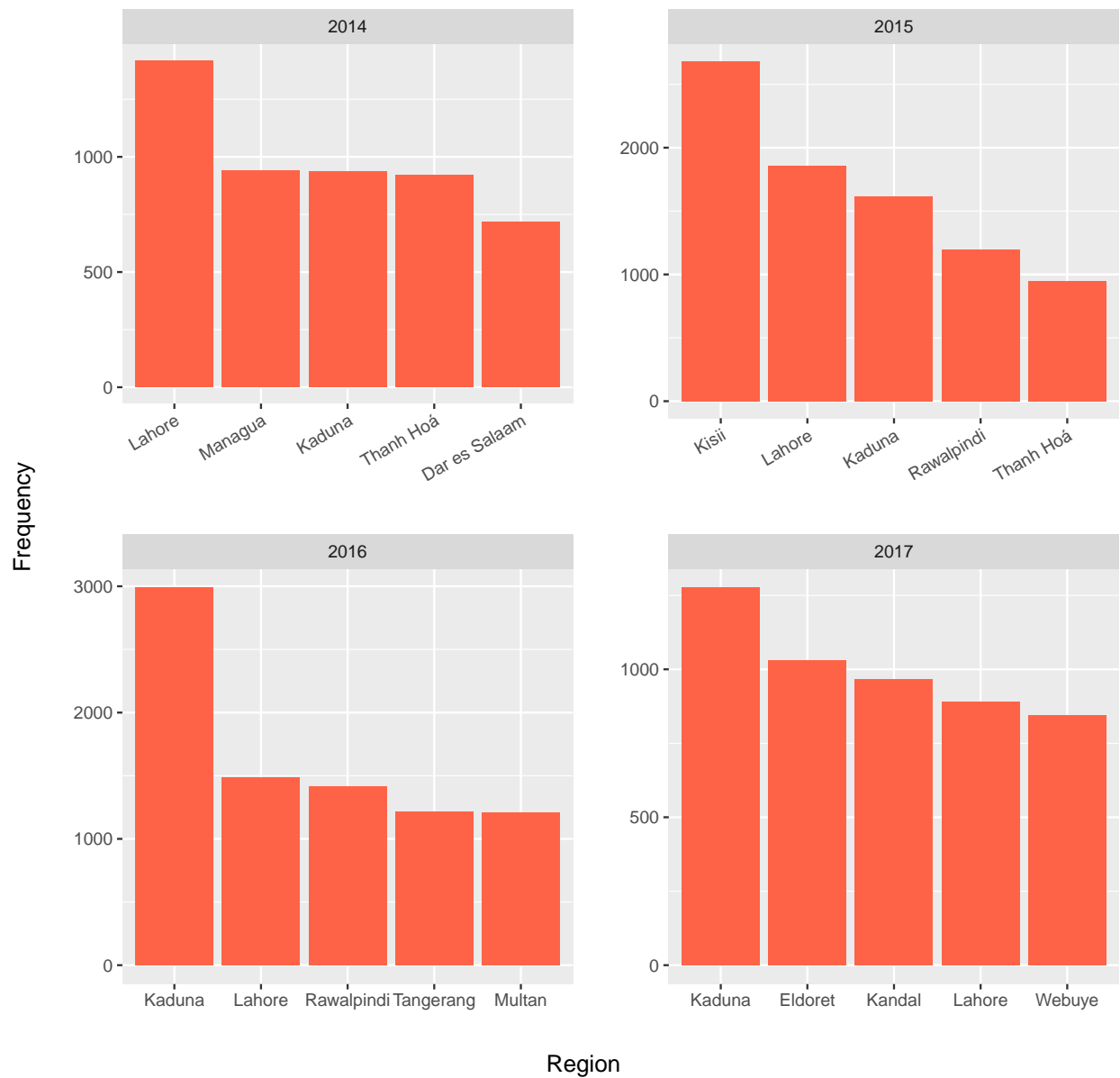
From the bar plot, there is a mix of top regions that are top funded per year.

Is there a relationship between frequently funded regions/countries and *MPI*?

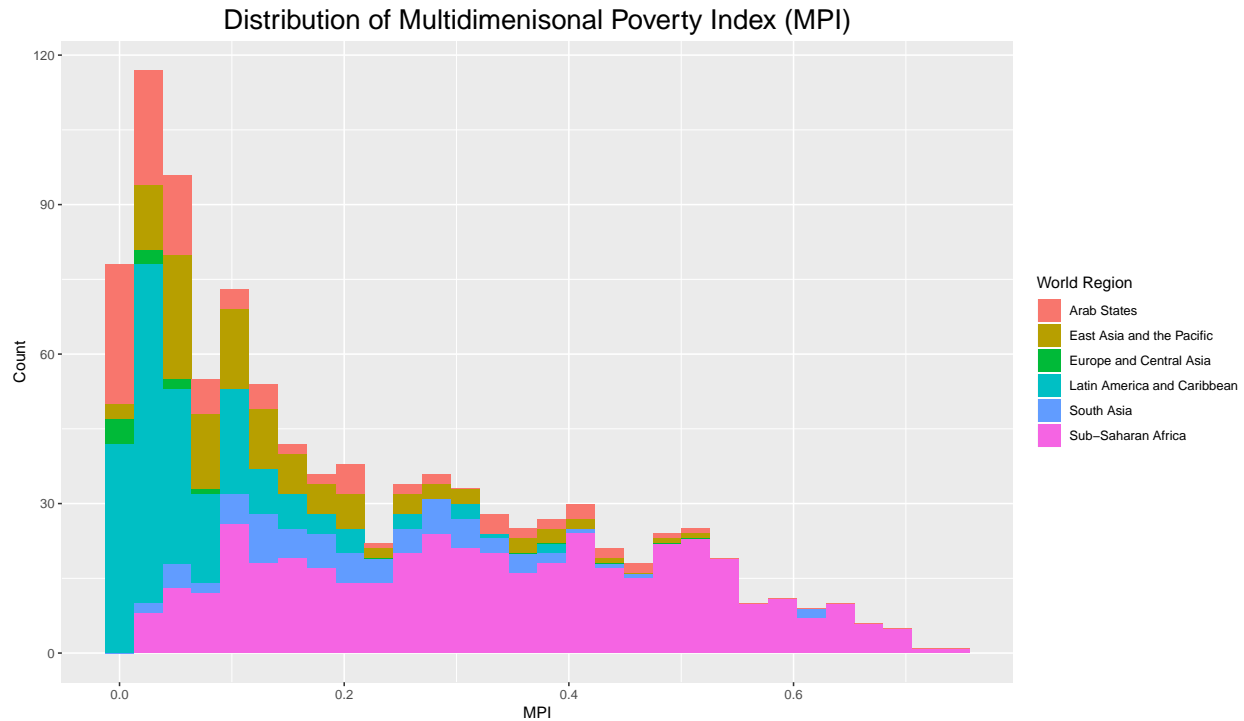## Funded Loan Amount and Poverty Index (MPI)

Here we introduce MPI and the 6 World Regions. This includes only regions and countries with an MPI.

```
## [1] "Max MPI = 0.74"
```

```
## [1] "Min MPI = 0.00"
```

```
## [1] "Med MPI = 0.15"
```
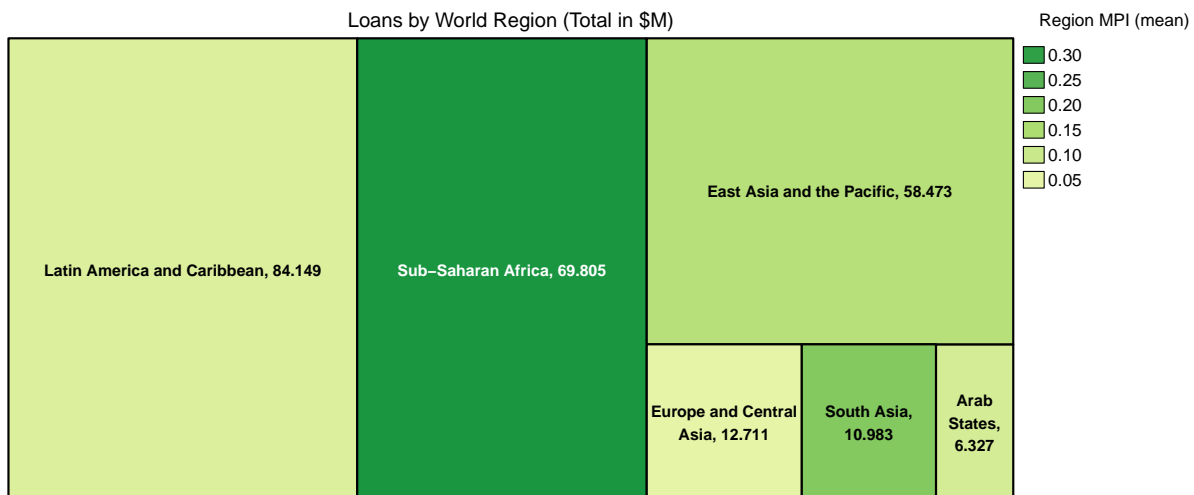
```
## [1] "Mean MPI = 0.21"
```

Distribution of Multidimenisonal Poverty Index (MPI)

From distribution, the poorest world regions are Sub-Saharan Africa and South Asia. What proportion of loans are these *WorldRegion*s receiving?

```
## [1] "No. of Total Regions = 928"
```

```
## [1] "No. of Total Countries = 102"
```
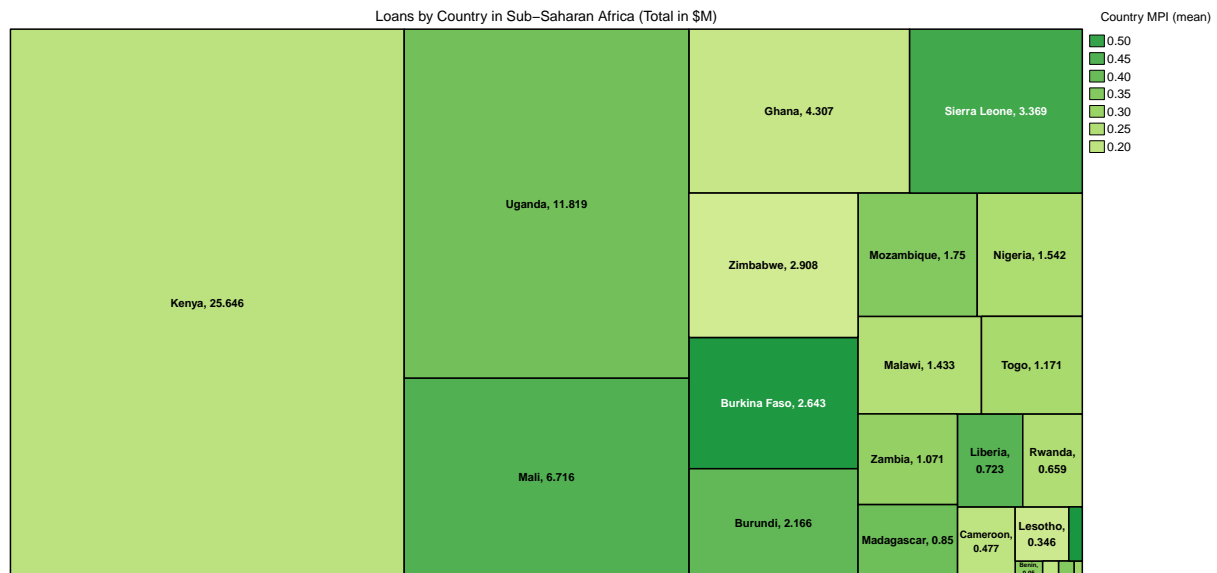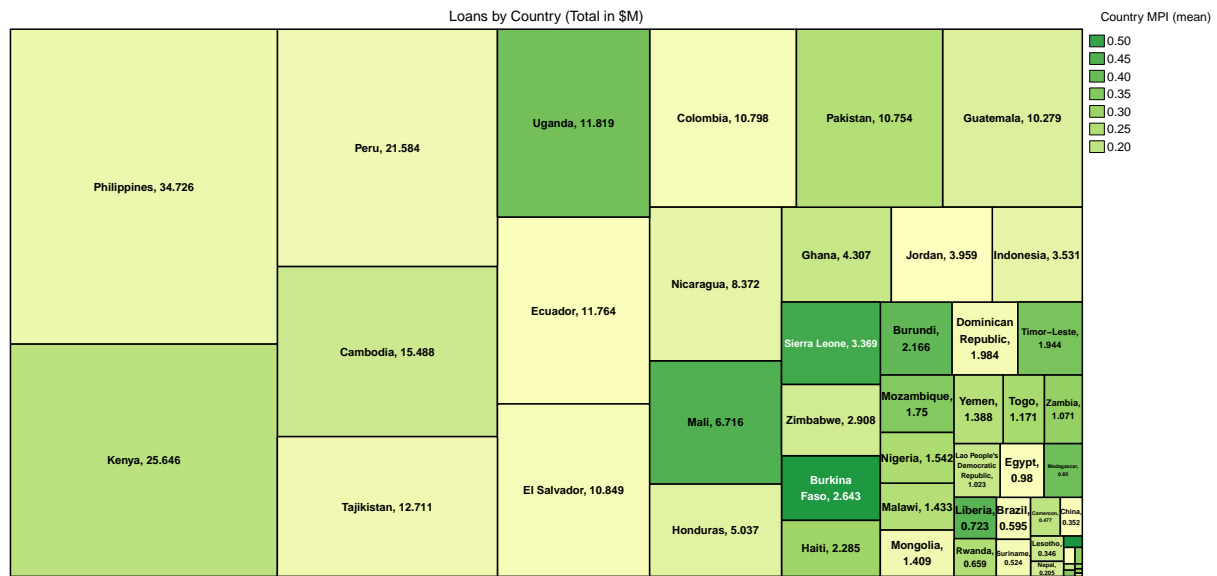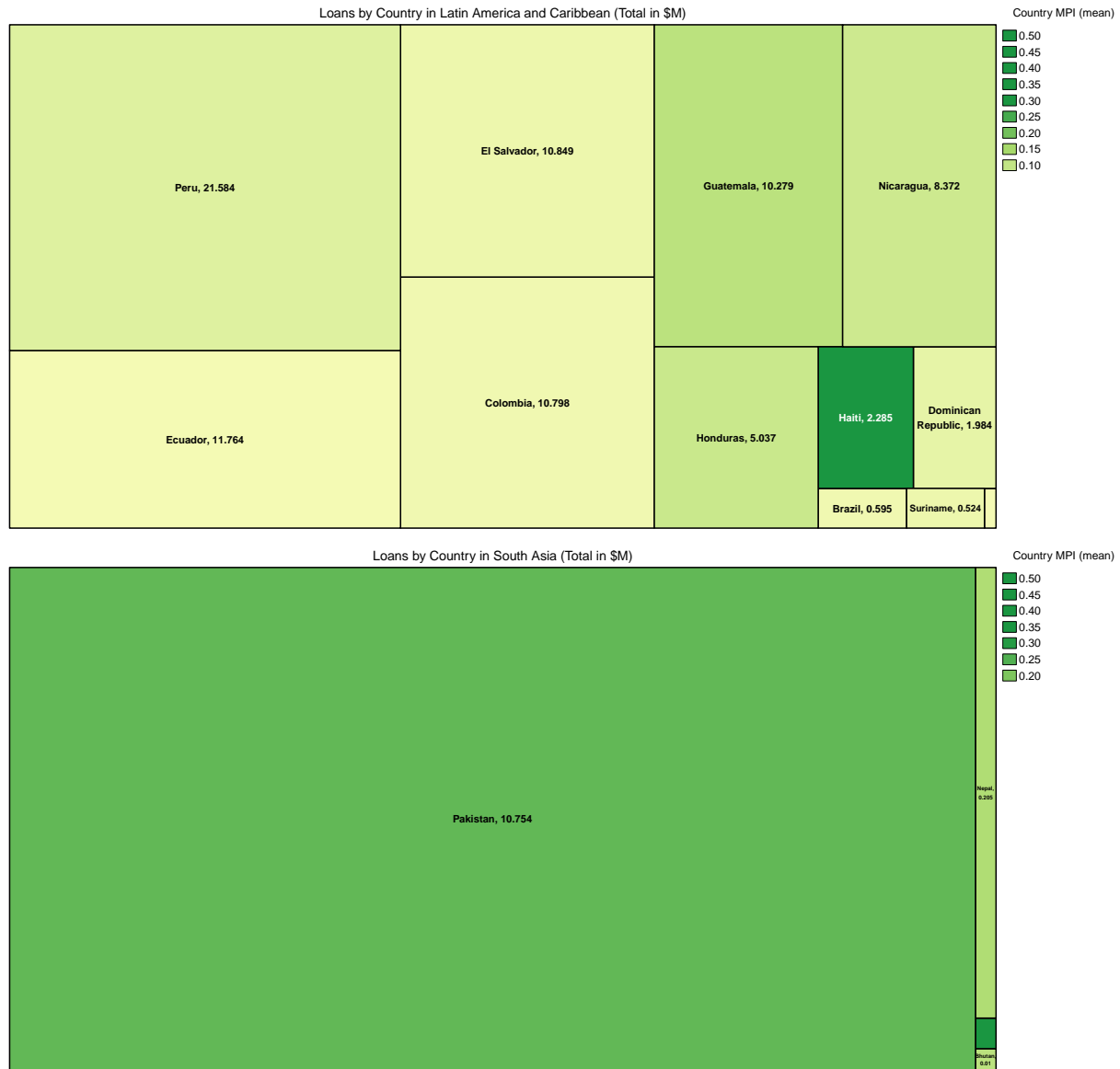
```
## [1] "No. of Total World Regions = 6"
```


As we saw from the *MPI* distribution, Sub-Saharan Africa is the poorest *WorldRegion*. From the treemap, Sub-Saharan Africa has received a large portion of the total funded loans. While South

Asia, high on the poverty index, receives the second smallest portion of funded loans. South Asia might be an area to focus on to identify loan trends.

Loans by Country (Total in $M)

Country MPI (mean)
- 0.50
- 0.45
- 0.40
- 0.35
- 0.30
- 0.25
- 0.20

Philippines, 34.726
Peru, 21.584
Uganda, 11.819
Colombia, 10.798
Pakistan, 10.754
Guatemala, 10.279
Cambodia, 15.488
Ecuador, 11.764
Nicaragua, 8.372
Ghana, 4.307
Jordan, 3.959
Indonesia, 3.531
Kenya, 25.646
Tajikistan, 12.711
El Salvador, 10.849
Mali, 6.716
Sierra Leone, 3.369
Burundi, 2.166
Dominican Republic, 1.984
Timor–Leste, 1.944
Zimbabwe, 2.908
Mozambique, 1.75
Yemen, 1.388
Togo, 1.171
Zambia, 1.071
Nigeria, 1.542
Lao People's Democratic Republic, 1.023
Egypt, 0.98
Madagascar, 0.85
Burkina Faso, 2.643
Honduras, 5.037
Malawi, 1.433
Liberia, 0.723
Brazil, 0.595
Cameroon, 0.477
China, 0.352
Haiti, 2.285
Mongolia, 1.409
Rwanda, 0.659
Suriname, 0.524
Lesotho, 0.346
Nepal, 0.285

Loans by Country in Sub–Saharan Africa (Total in $M)

Country MPI (mean)
- 0.50
- 0.45
- 0.40
- 0.35
- 0.30
- 0.25
- 0.20

Kenya, 25.646
Uganda, 11.819
Ghana, 4.307
Sierra Leone, 3.369
Zimbabwe, 2.908
Mozambique, 1.75
Nigeria, 1.542
Burkina Faso, 2.643
Malawi, 1.433
Togo, 1.171
Mali, 6.716
Zambia, 1.071
Liberia, 0.723
Rwanda, 0.659
Burundi, 2.166
Madagascar, 0.85
Cameroon, 0.477
Lesotho, 0.346
Benin, 0.05

Loans by Country in Latin America and Caribbean (Total in $M)

Loans by Country in South Asia (Total in $M)

From these treemaps, some of the poorest countries are receiving a small portion of the total Kiva loans. Can see the darker green more prominent in the lower most corner.

Burkina and South Sudan within Sub-Saharan Africa, Haiti within Latin America and Caribbean, and Afghanistan within South Asia are receiving a small portion of the funded amounts.

Again, these areas of focus for loan trends and what is driving these differences. One thing to consider are potential sector and activity differences between the regions/countries. Do the loan needs of the poorer countries cost less than the others? Can we use this to estimate poverty levels and needs for those countries?

Frequently Funded Sectors

Top 5 Funded Sectors per Year
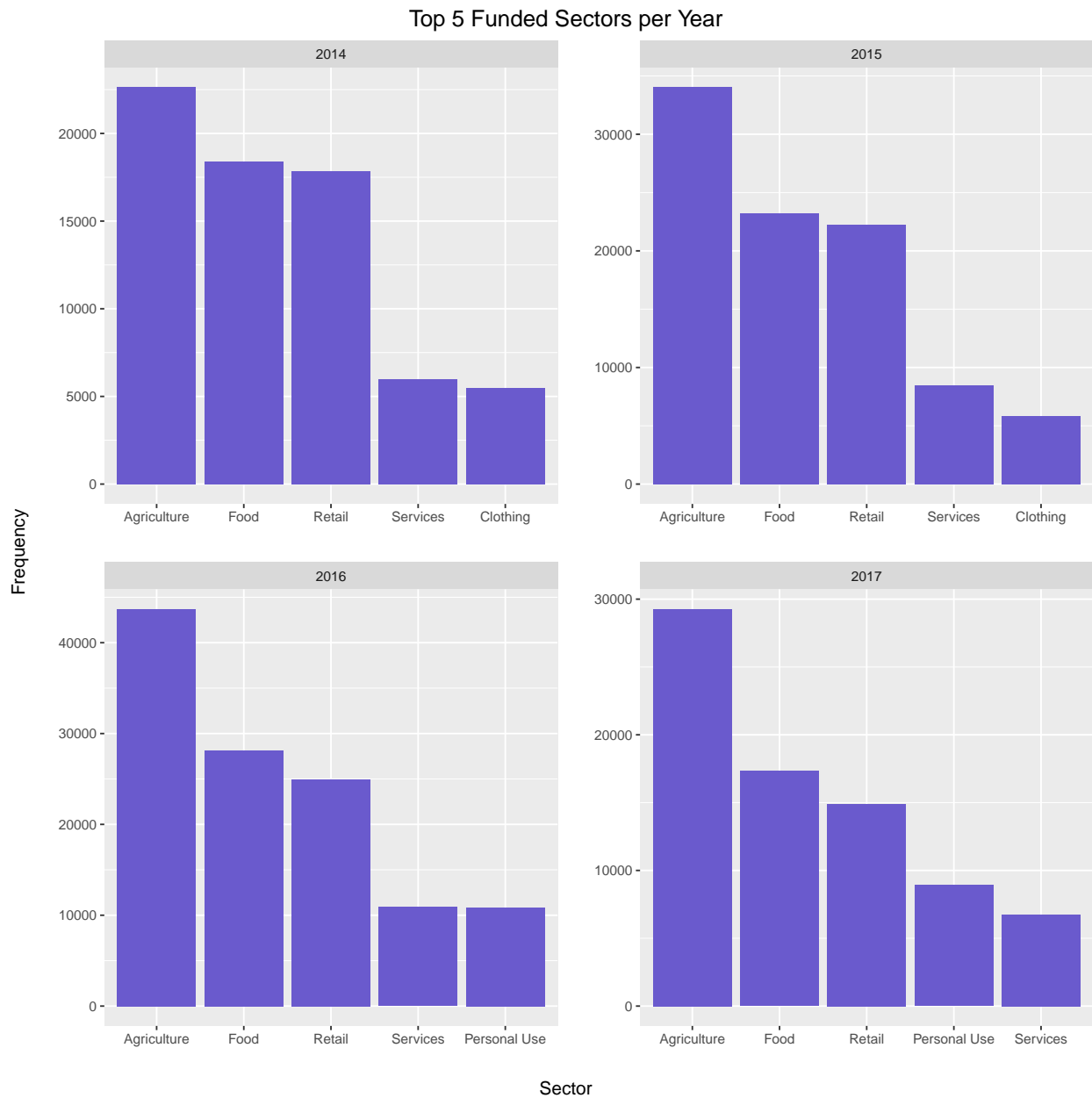
From the bar plots, the most frequently funded *sector*s are consistently Agriculture, Food, and Retail. What is the overall loan distribution among the sectors?

Distribution of Funded Loans by Sector (Log Scale)
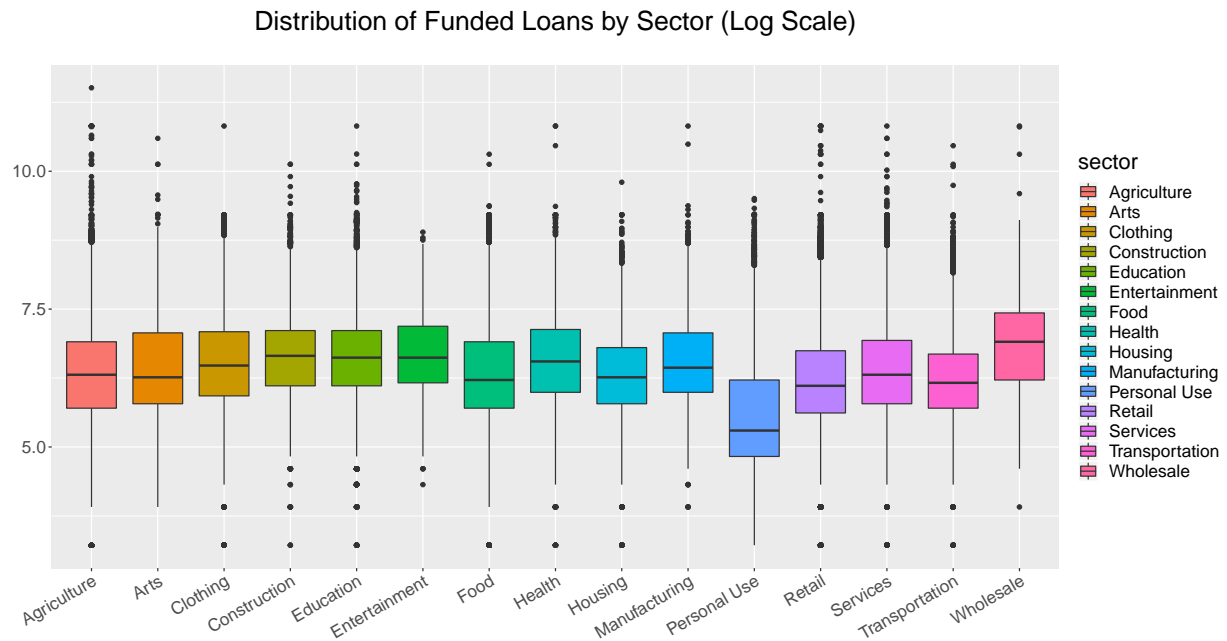
From the box plot, there is some variation in the loan amounts among the *sector*s. The Food, Housing, and Personal Use sectors have the lowest medians. This could correspond to what we saw from the treemaps. So, what is the funded loan breakdown for these top sectors for World Region and Country?

Proportion of Funded Loans for Sectors by World Region

```
## Warning: Column `country` joining factors with different levels, coercing
## to character vector
```

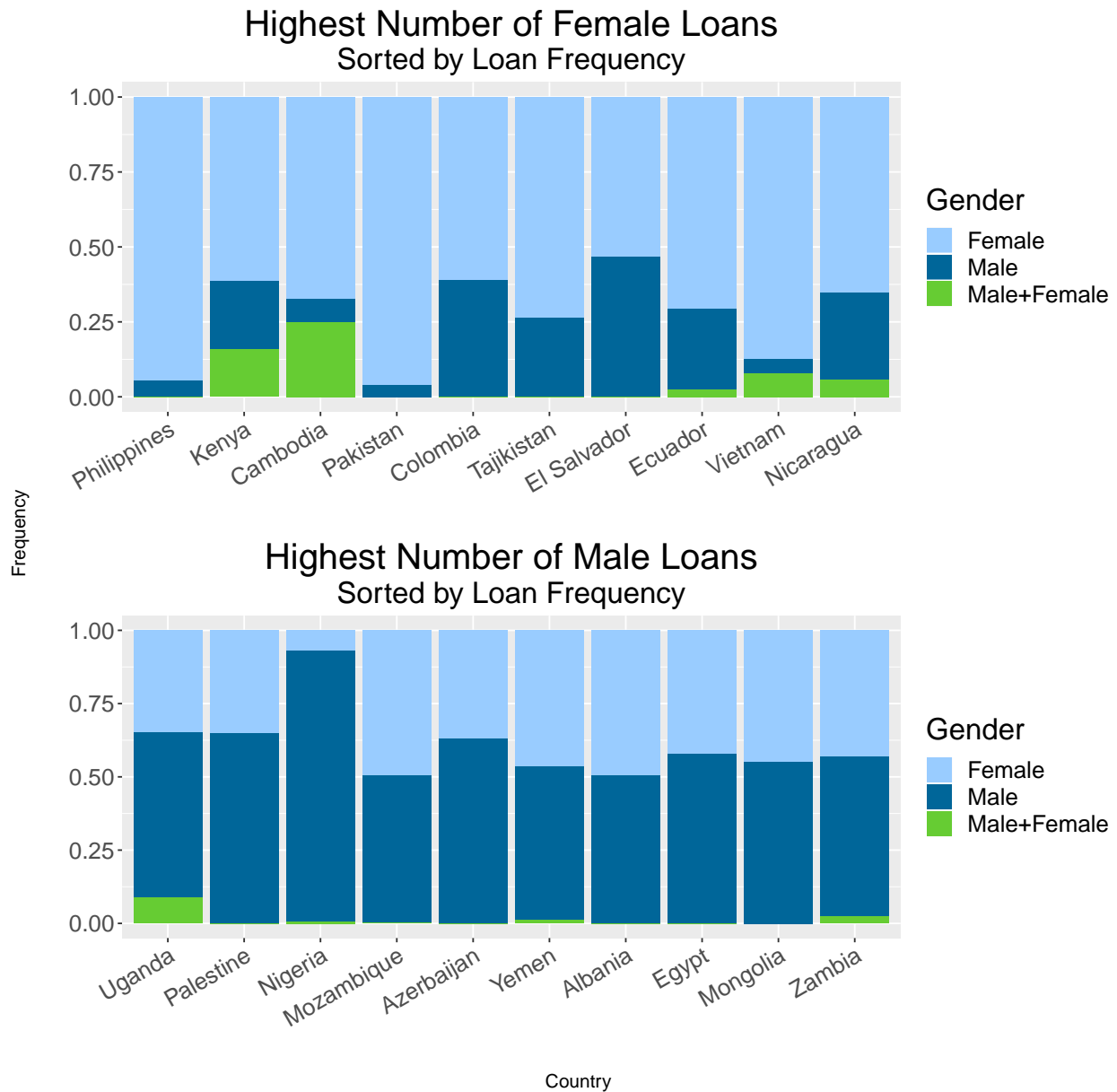Relationship Between World Regions and Top Sectors
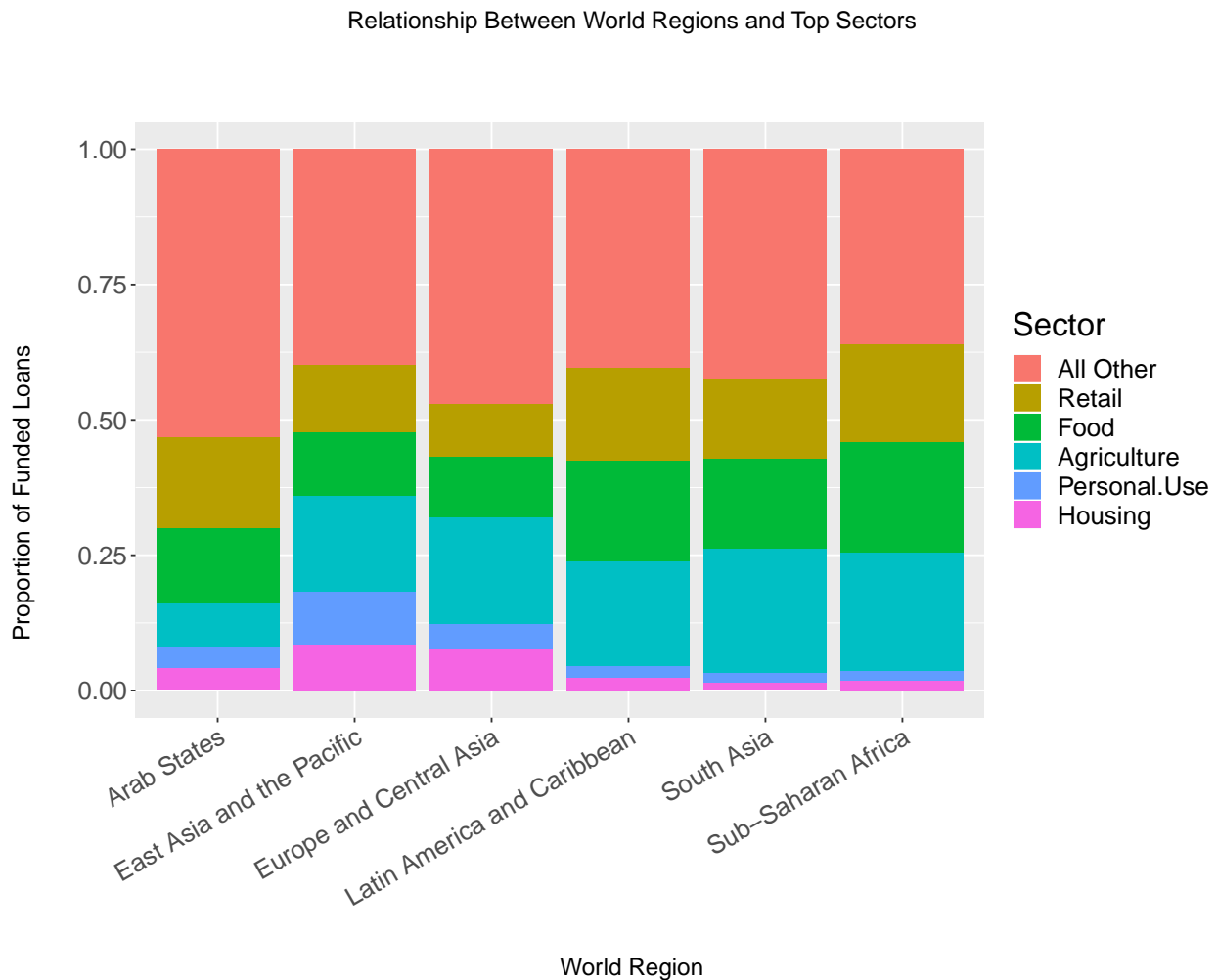
The chart indicates that Agricultural loans make up a good portion of the funded loan amounts. Recalling both the treemap by World Region and the sector boxplots, I would have expected a larger proportion of Personal Use and Housing within Sub-Saharan Africa and South Asia. (I will revisit this). Now we will take a look how *Gender* plays a role in the dataset.
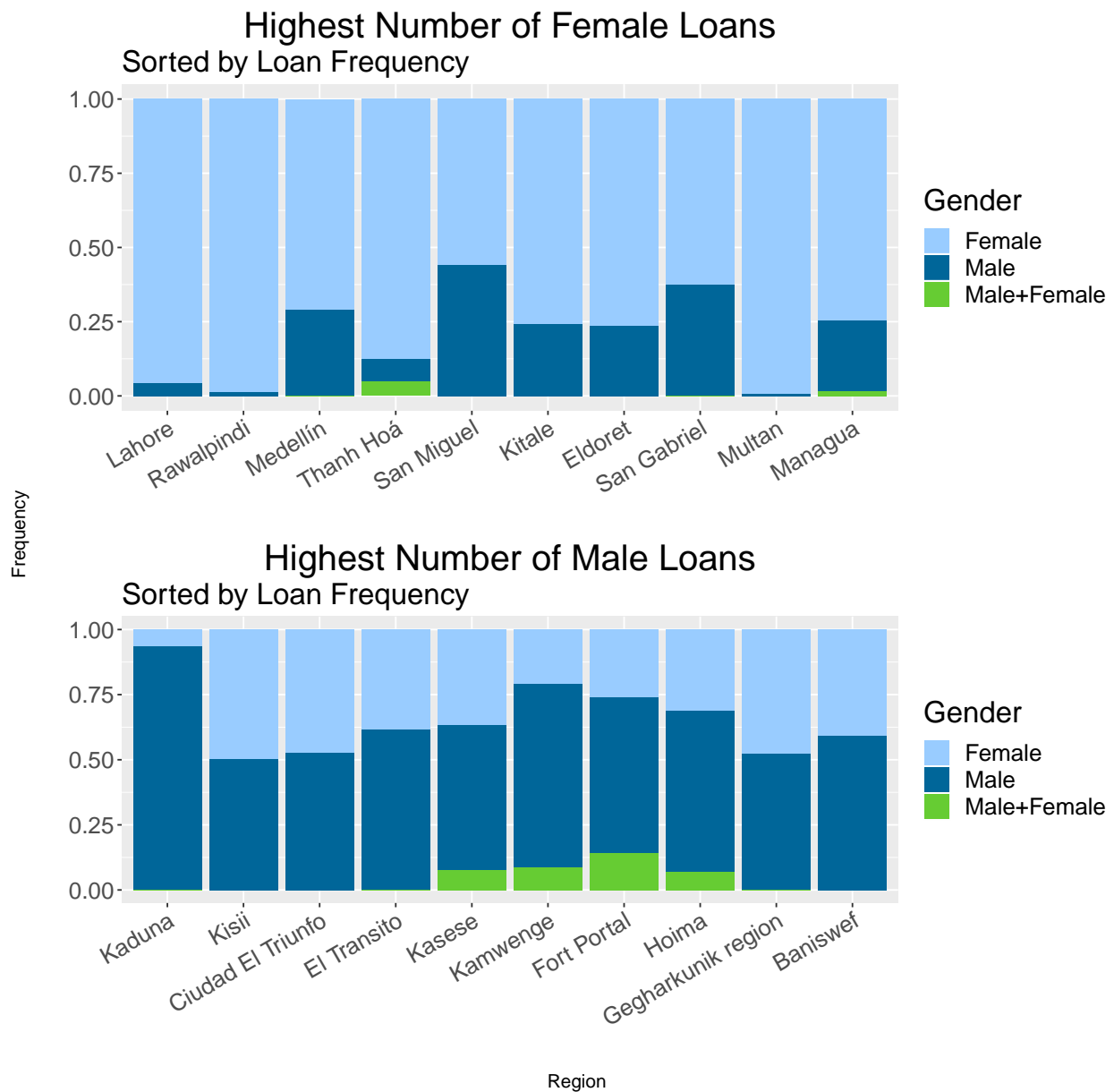
Loan Breakdown by Gender

Now that we have a sense for the loans

```
## [1] "No. of Total Loans = 462167"
```

```
## [1] "Female Only Loans = 70%"
```

```
## [1] "Male Only Loans = 23%"
```

```
## [1] "Female+Male Loans = 7%"
```

```
## [1] "Total Loans - Funded = 92%"
```

```
## [1] "Female Only Loans - Funded = 94%"
```

```
## [1] "Male Only Loans - Funded = 83%"
```

13

[1] "Female+Male Loans - Funded = 92%"

Highest Number of Female Loans
Sorted by Loan Frequency

Highest Number of Male Loans
Sorted by Loan Frequency

Frequency

Country

Highest Number of Female Loans
Sorted by Loan Frequency

Highest Number of Male Loans
Sorted by Loan Frequency

These charts provide good summaries for the *Gender* differences across countries and regions. There are clear differences among the countries and regions for who is taking out the loan.
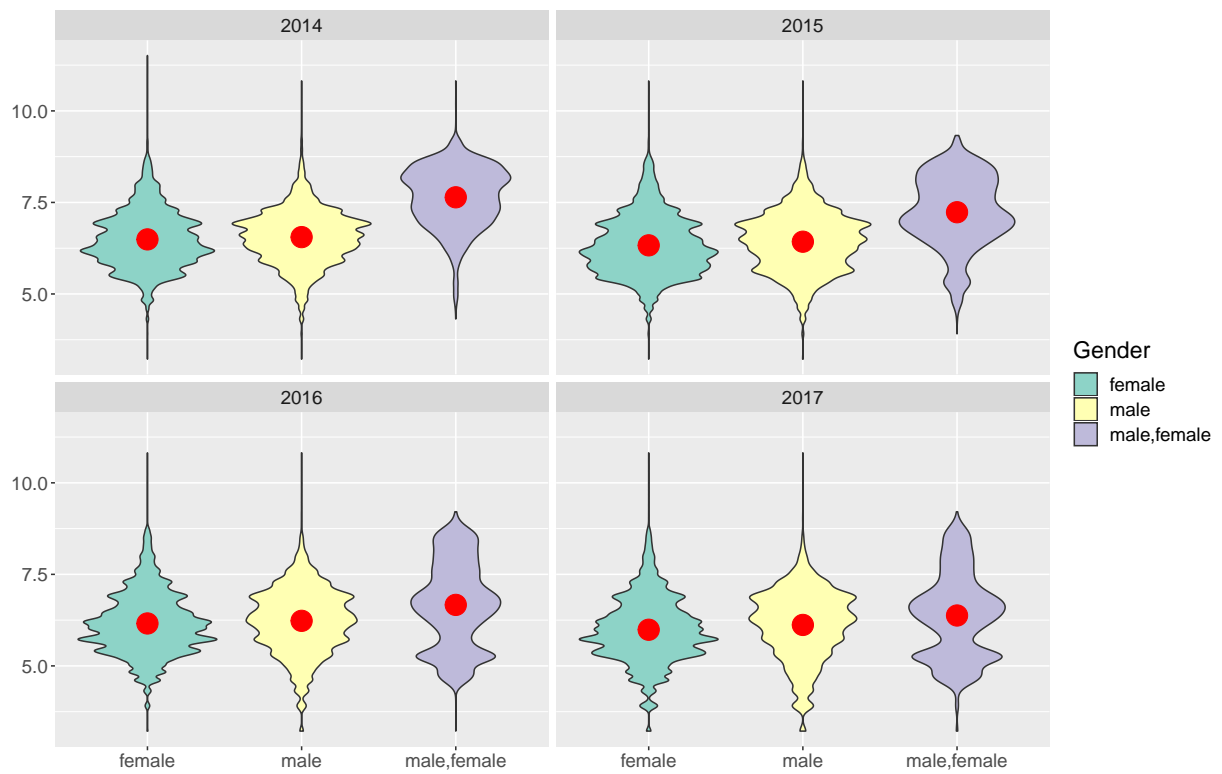
Loan Breakdown by Gender

Relationship Between Funded Loan Amount and Gender
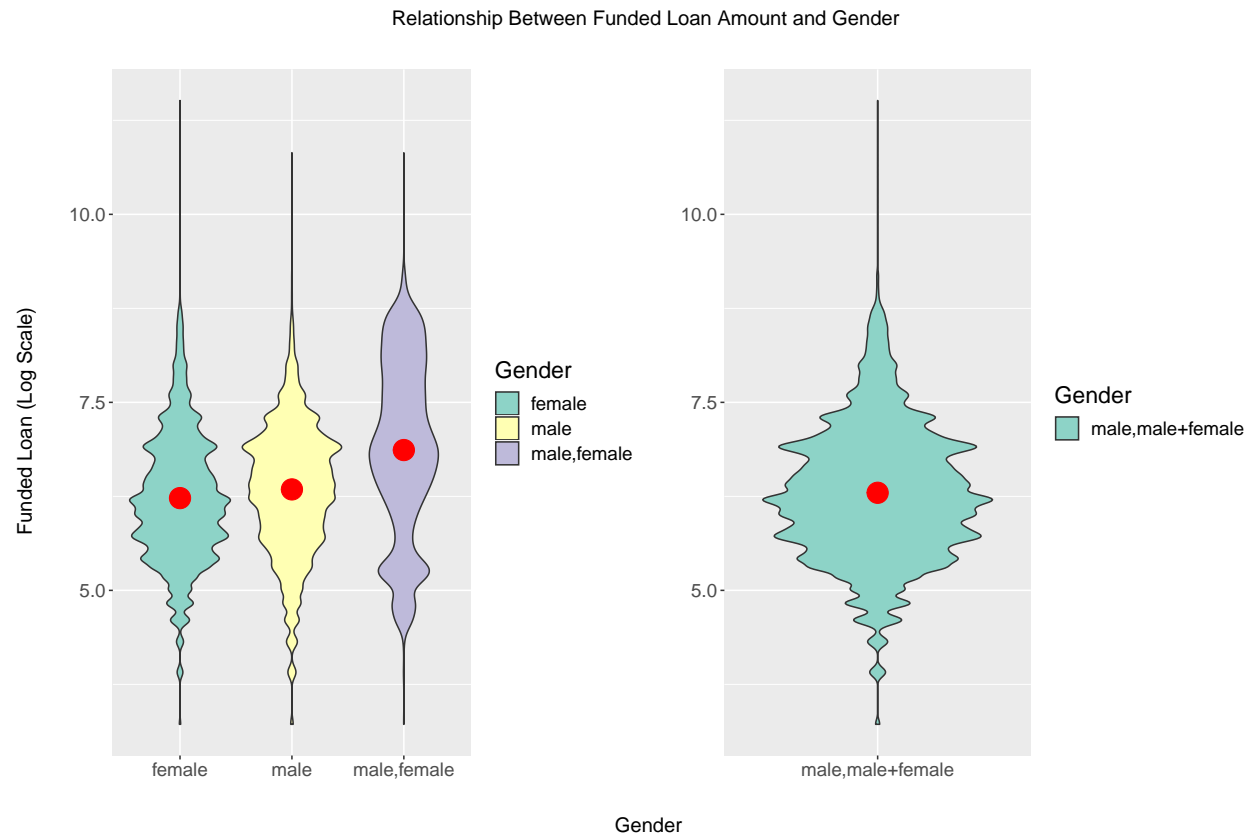
From the violin plots, there is a difference in the average funded loan amount (red dot) between females and males on an overall basis.
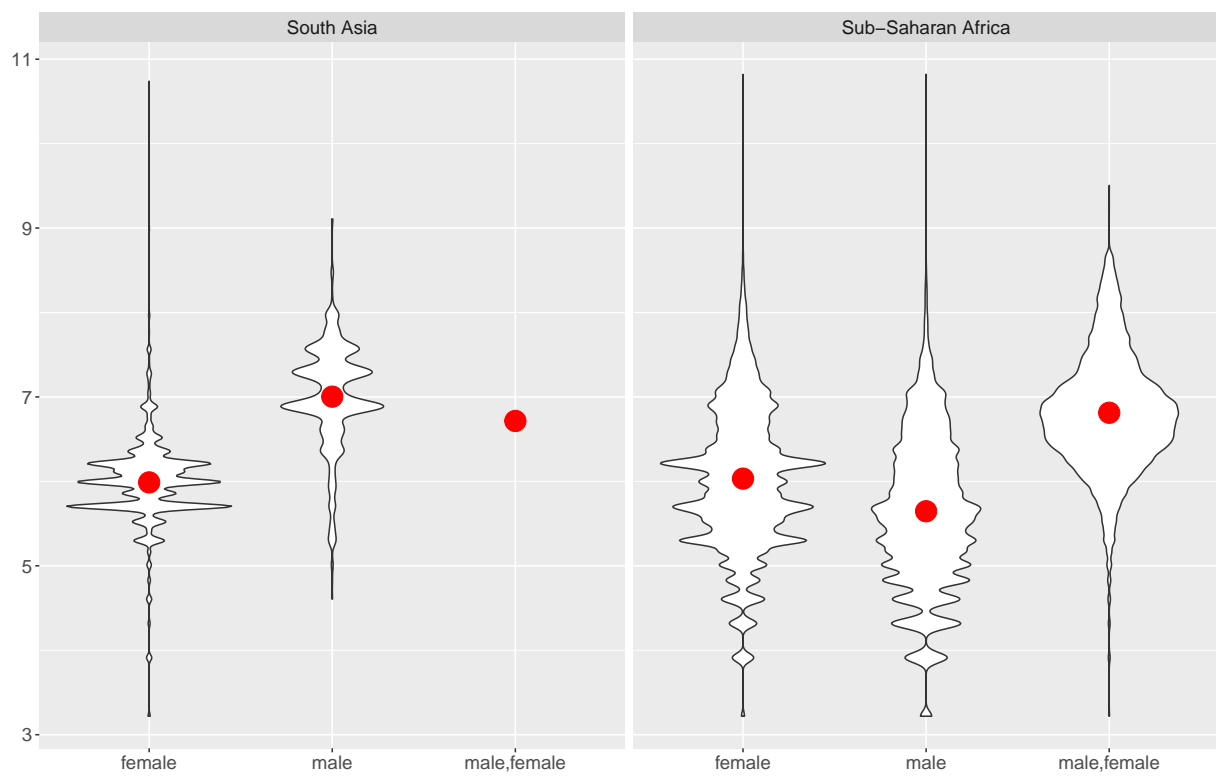
We may see even bigger differences by looking at the gender differences across World Regions and Countries and over time.
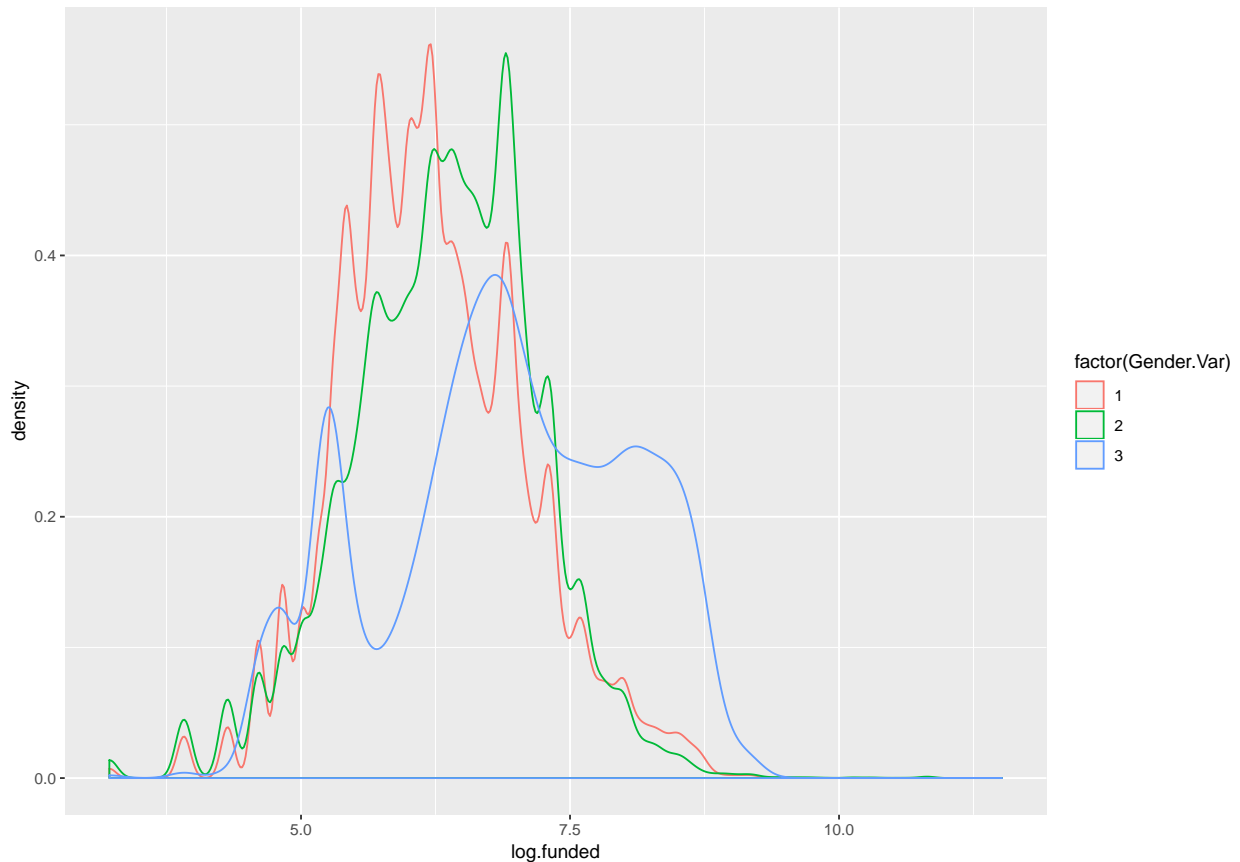
From the violin plots by year, the average loans overall seem to be decreasing, but also leveling between female, male, and male+female.

We may see something interesting across World Regions and Countries.

There was just one loan that had male,female variable for South Asia (in 2014). (We will revisit this in more detail in our modeling)

Multilevel Model

This is a work in progress. There are many different levels to this dataset.

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log.funded ~ Gender.Var + sector + (1 + Gender.Var | country)
##    Data: fit1.data
##
## REML criterion at convergence: 924421
##
## Scaled residuals:
##     Min       1Q  Median       3Q      Max
## -7.0133 -0.5798 -0.0108   0.6249   8.4946
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  country  (Intercept) 1.2636   1.1241
##           Gender.Var  0.1395   0.3734   -0.77
##  Residual             0.4356   0.6600
## Number of obs: 460113, groups:  country, 82
##
## Fixed effects:
```

```
##                    Estimate  Std. Error   t value
## (Intercept)         6.790185   0.128360    52.899
## Gender.Var          0.130558   0.044362     2.943
## sectorArts          0.069592   0.007852     8.863
## sectorClothing      0.070697   0.004977    14.204
## sectorConstruction  0.079784   0.010421     7.656
## sectorEducation    -0.081100   0.004993   -16.241
## sectorEntertainment 0.167618   0.030477     5.500
## sectorFood          0.062365   0.003119    19.993
## sectorHealth       -0.126373   0.008485   -14.894
## sectorHousing      -0.206266   0.005106   -40.400
## sectorManufacturing 0.159284   0.010891    14.626
## sectorPersonal Use -0.820076   0.005366  -152.837
## sectorRetail        0.067672   0.003238    20.897
## sectorServices      0.026327   0.004409     5.971
## sectorTransportation -0.043479 0.006771    -6.421
## sectorWholesale     0.381844   0.030321    12.594
```

Conclusion/Next Steps

There is a lot to consider in this analysis. This project is ongoing and will be diving deeper into modeling next. It is important to understand the underlying themes and behaviors that differ between regions and countries. This data can help Kiva in supporting these areas.
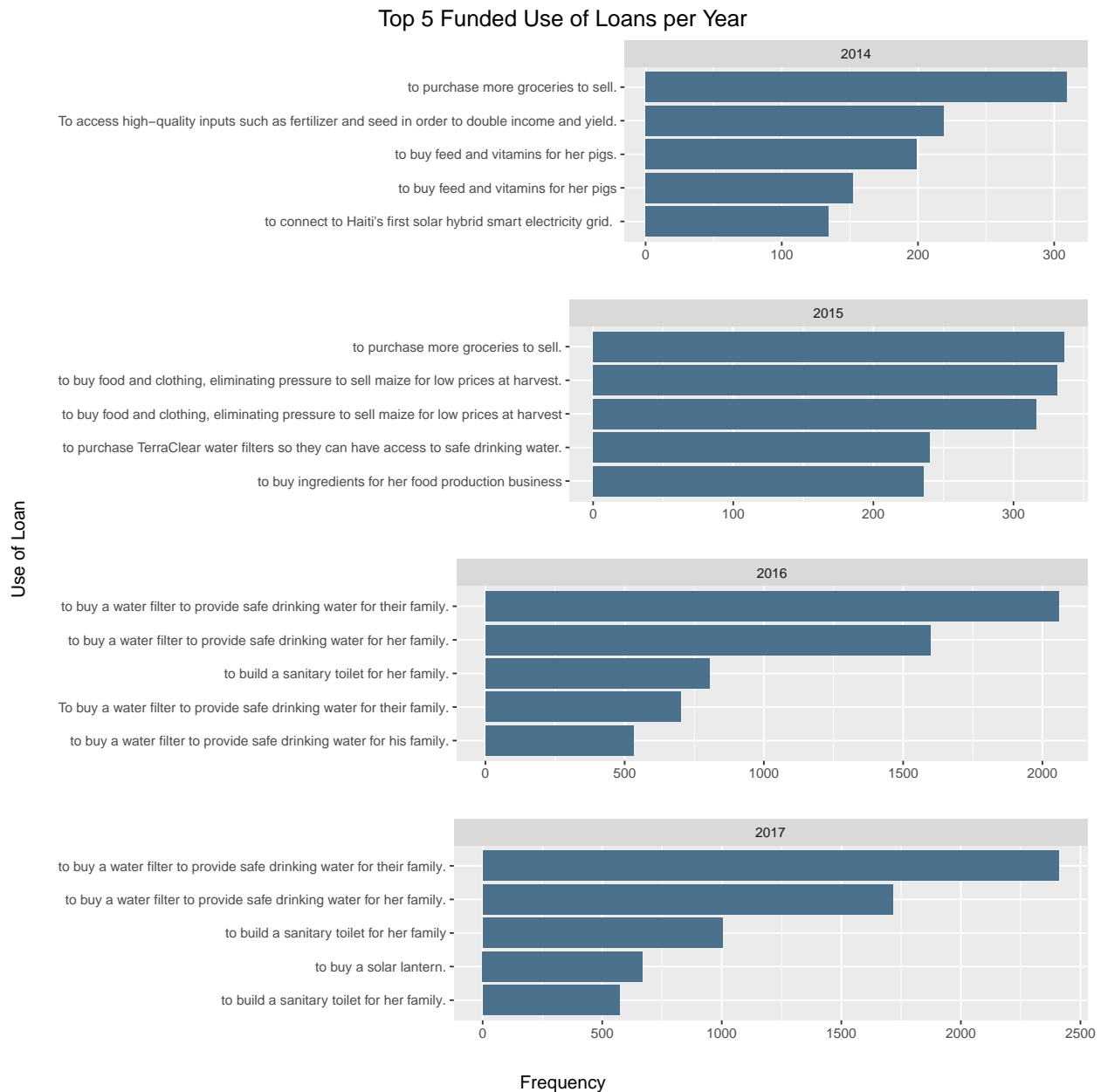
Appendix

Items to look at in the future:

Where is the use of water loan most prevelant?

Top 5 Funded Water Loans

How are loans being used?

Top 5 Funded Use of Loans per Year
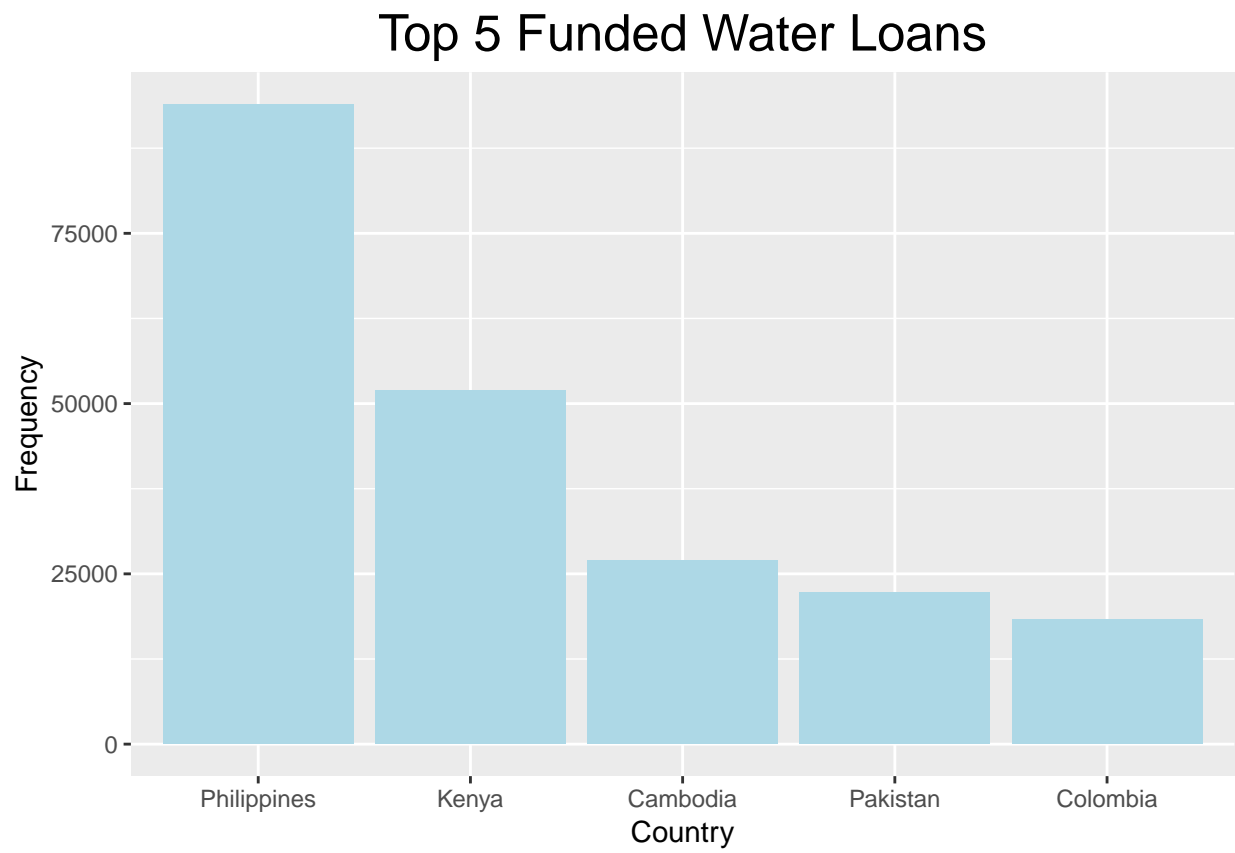
From the bar plots, there is some variation in the use of the loan. For water uses, does this vary by region and become more prominent during dry seasons? If there is an expected dry season can we expect water loans to increase?

```
#- Most imporverished areas
#- What is being funded/partially funded/not funded and likelihood?
#- Female vs. male borrowers and whether having a male in the group affects loan behavior?
#- Repeat and type of borrowers
#EDA plot outcome on number of F and number of male and ratio of male count to female count to i
#Does a male impact on the loan.
```

```
#Number of females might not matter, but once adding in a male that could affect the loan amount
#Linear regression model per country on amount of loan for gender
```