Capstone Project I

- **Problem statement and motivation:**

The client is Kiva.org. The goal is to understand what may cause a higher delay between the 'posted time' (the time at which the loan is posted on Kiva by the field agent), the 'funded time' (the time at which the loan posted to Kiva gets 100% funded by lenders), and the 'disbursed time' (the time at which the loan is disbursed by the field agent to the borrower). Subsequently, have a sense of its urgency and possibly discover hints on how they could resolve the issue. The problem is important since this could be a way of understanding a fundamental step in its business process, and take actions wherever needed. Lenders would get their borrowed money faster and it could possibly be an incentive to lend again more times, and/or with more money. Depending on what insights the analysis of the data provides, Kiva can then think about what they will do.

1) Data inspection and cleaning steps

After getting an overview of the data, I **changed some column names, for clarity, and changed the dtype of the main variables to datetime64[ns].**

1.1) Main variables inspection and cleaning

'Posted_time'

- There were no missing values and the filled ones seemed to be ok - Apparently, the 'posted_time' column is clean.

'Funded_time'

There are 64282 missing values and 1355325 filled values.

- **Regarding the missing values:**
    - Even though some null-values might correspond to loans which were indeed funded, there is not a way to know how to fill those values. Therefore, I **removed the rows where there is a missing value in funded_time but (loan_amount - funded_amount) <= 0,** so that there is no ambiguity. Note: I considered ok when the funded_amount was larger than the loan_amount, hence the <).
    - The remaining missing values were kept due to several reasons. Not every loan is funded or when there is a filled value in funded_amount, because the loan may not be completely funded yet, that is, the funded_amount is smaller than the loan_amount.

- **Regarding the filled values**, I compared with two columns to check if there were discrepancies: funded_amount and status.
    - *funded_time vs funded_amount:*
        - Comparing with the funded_amount column, there were no discrepancies.
        - However, when looking at the difference between loan_amount and funded_amount, there was inconsistency in the data. I then **removed the rows where the funded_time was filled, but (loan_amount - funded_amount) > 0**. Note: the funded_time is only filled when the loan is 100% funded by lenders.
    - *funded_time vs status:*
        - The status column is composed of 4 hypotheses: funded, refunded, expired and fundraising.
        - Comparing with the *funded* status, everything was ok.
        - Comparing with the *refunded* status, everything was ok.
        - Comparing with the *expired* status, there was inconsistency in the data. There were two loans requests which supposedly expired but, at the same time, two values for the 'funded_time' column were filled. In this case, however, the remaining variables seemed to indicate that the loan was indeed funded, so I considered them for analysis. For example, in the first case, what probably classified the status as expired was the fact that the funded_time was slightly after (6 minutes) the planned_expiration_time. I **changed the status in these two cases to 'funded'.**
        - Comparing with the *fundraising* status, everything was ok.

'Disbursed_time'
The "disbursed time" is the time at which the loan is disbursed by the field agent or group of lenders to the borrower.

- **Regarding the missing values:**
    - It may be the case that there was a posted_time or funded_time but the disbursal has not occurred yet. I will leave those values for further analysis since they could represent loans which were not delivered or just expected situations.
- **Regarding the filled values:**
    - The timing of the disbursal can vary. For most Field Partner loans, the money is pre-disbursed, so the borrower can access the funds right away. Hence, it is not strange if the disbursed_time is made even before the posted_time. For direct loans, the money is disbursed only after the loan has been fully crowdfunded on the Kiva website. The disbursed_time can then naturally occur after or before the posted_time or the funded_time. All of the values remained for analysis.

'Posted_time' vs 'funded_time'
The chronological order between the 'posted_time' and the 'disbursal_time' does not matter, as well as between the 'disbursal_time' and the 'funded_time'. What could bias the data is when the

'funded_time' appears before the 'posted_time'. There were indeed found cases where the funded_time was way before the posted_time, and they all corresponded to the first posting date. Therefore, I **removed these cases** for analysis.

1.2) Other variables

- The funded_amount column was compared with the num_lenders_total column and everything was ok.
- If the funded_amount is greater or equal to the loan_amount, then the status should not appear as expired. No inconsistency was found.
- Checking for misspellings: No errors were found in the activity_name and country_name columns.

1.3) Searching for outliers

'loan_amount' and 'funded_amount'

In the previous analysis, when looking at the filled values in the 'funded_time' column, I did not include the cases where the funded_amount was greater than the loan_amount because it is acceptable. I checked, however, for outliers. Just looking at the general statistics of the differences, it was possible to see that they were not significant. Therefore, I **kept all of the data**.

Main variables

As the main variables fit approximately inside the same dates, there are no outliers regarding each one. The next step was to look at the differences between them.

'posted_time' and 'funded_time'

It was not necessary to look for outliers here since I had already excluded for analysis the cases where the funded_time was before the posted_time. The rest of the values **were accepted** at face value since they are between an acceptable range.

'posted_time' and 'disbursed_time'

The chronological order of these two dates does not matter, because the disbursal can be made before the posted time for most of the entities (Field partners). It was necessary to check for outliers, though. I focused on the cases when the disbursal was made before the posted_time.

- I defined a z-score function to detect outliers.
- With a threshold of 3, a few outliers were found. However, when examining them, in all of the cases the difference between the posted_time and funded_time was acceptable/minimal, which possibly means that although the disbursal was made long before the posted_time, the loan got funded almost immediately. This could suggest that these cases were somehow managed by Kiva or the Field Partners. I then **opted for their maintenance** in the analysis. (A box plot was also presented).

'funded_time' and 'disbursed_time'
Here I focused on the cases on the cases where the disbursed_time was filled after the loan was funded. I left apart from the cases where the disbursed_time appeared before the funded_time.

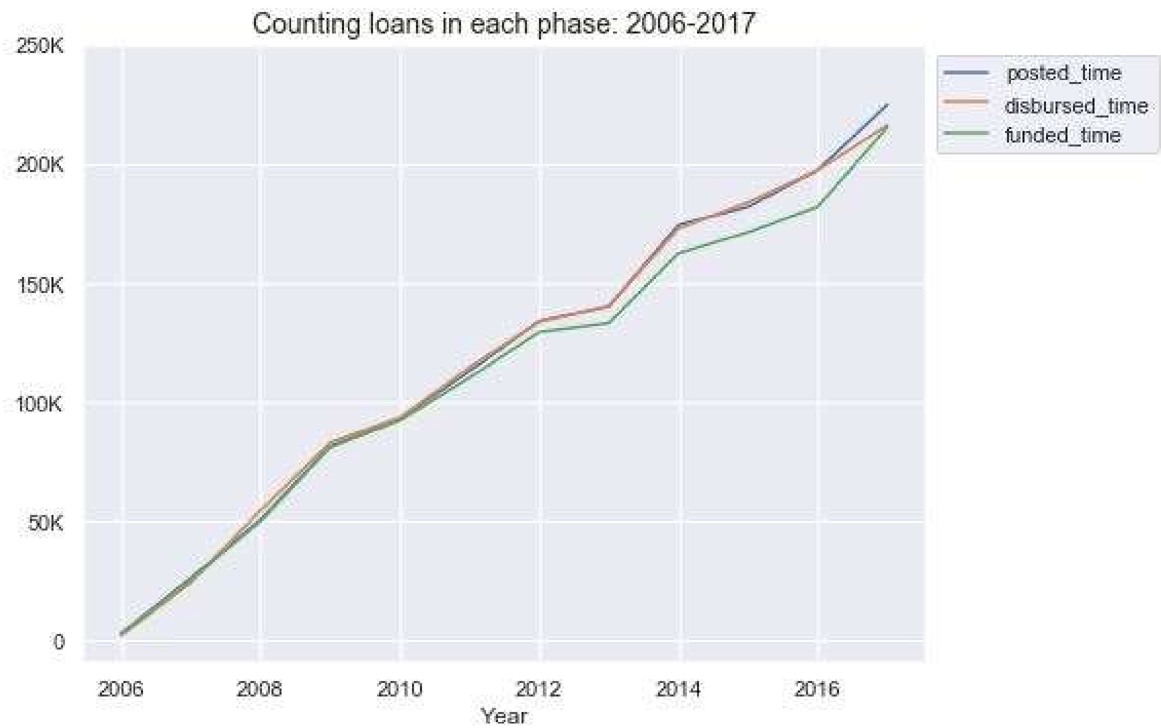
- After dividing the cases where there was a field partner involved and not, a few outliers were found, in both cases, using the previously defined z-score function with a threshold equal to 3. However, in both cases, when looking at the distribution of the outliers among countries, it was possible to see that they are somewhat representative, that is, they were concentrated in a small number of countries. I then **opted to keep the values** for further analysis.

There was also 271 missing disbursed_time values when the funded_time was filled, all of them of USA and Kenya, and the values were off the scale. I **removed them**.

2) Exploratory Data Analysis

Now that the data is clean, it was time to investigate the data, first using visuals. To do that, I performed the following steps:

I began to study the characteristics of the main variables and how they interacted with each other over time. Looking at the evolution of the number of loans in each phase, during the period 2006-2017, it is clear that there are three major distinctive time periods:



1. Until 2010, the amount of posted loans are closely matched in time by the number of fundings and disbursals.

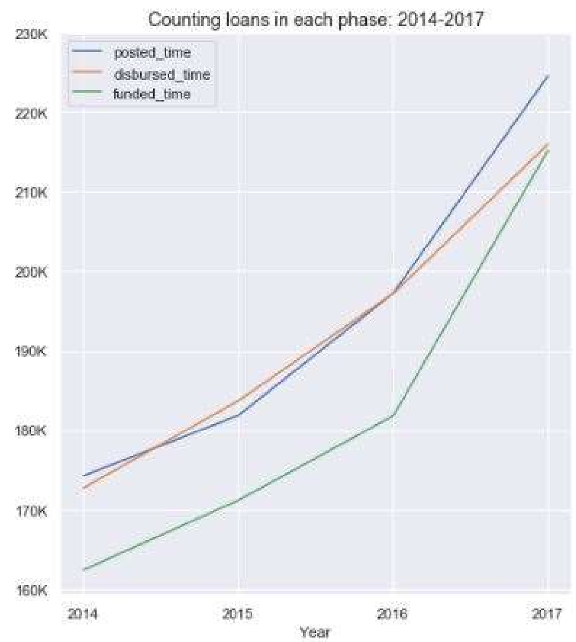
2. After that, between 2010 and 2014, there is a clear separation between each of the main variables: it suggests that the loans, continuing with almost always a close match between the posting phase and disbursal, started to lose its efficiency regarding the funding of the projects.

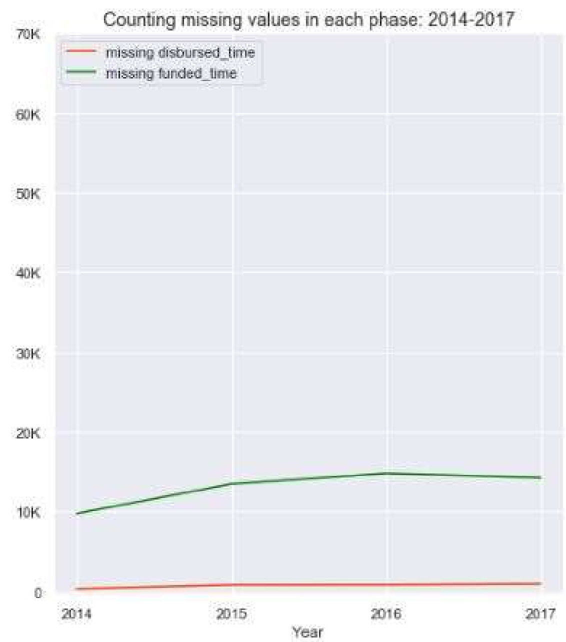
3. After 2014 till the end, the differences continued to aggravate between the postings and fundings, and the number of disbursals and postings became just a bit volatile over time. Curiously, the number of fundings gained proximity with the number of disbursals, especially in the last year.

This could be still be related to missing values or other factors, so it deserved further exploration.

The process was smooth until 2010. Between 2010 and 2014, the funding started to lose track of the posted loans mainly between 2010 and 2011. This difference, however, could be due to the cyclical nature of the process, as it was later observed. Thereafter, the difference was mostly due to the funded missing values.

In the last phase, although the difference between the number of missing values of the disbursals and fundings remained relatively constant, the number of fundings augmented more than disbursals, equalizing them on the final stretch.

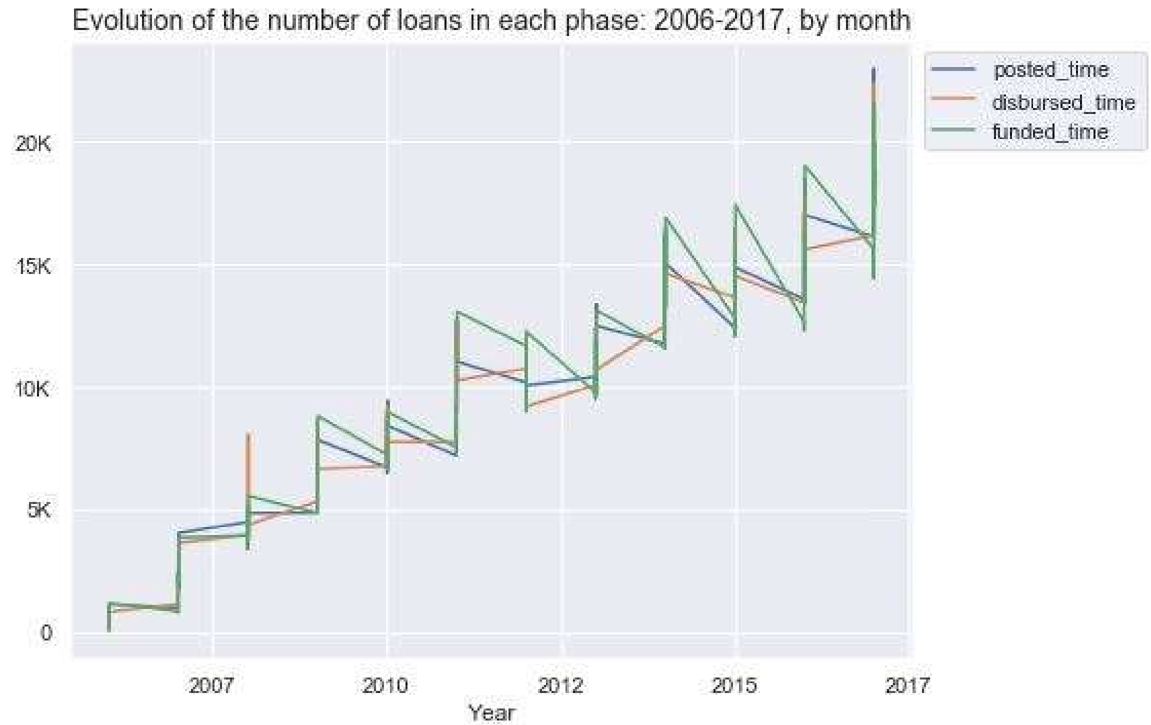




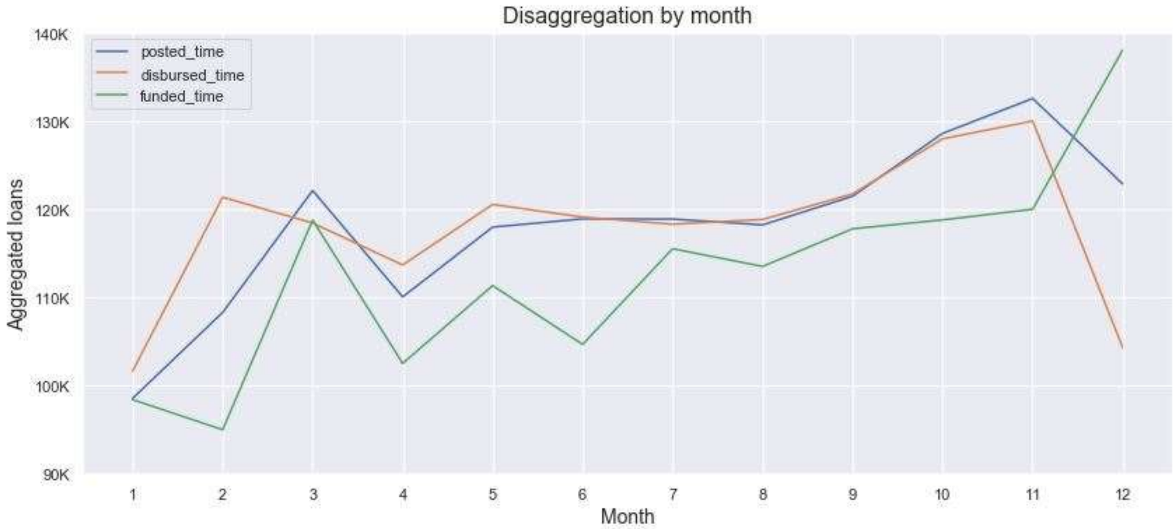
We can see that the deviation increase was mostly due to missing values until 2014. If we take that apart, the lines are somewhat at the same pace. This was not the case during the final 4 years, given the disbursals data we have. However, since it captures data at the end of the scale, and given the cyclical nature of the Kiva process, as shown below, it could still be related to the fact that the disbursal data was not yet inserted, lagging a few days or weeks.

In conclusion, we see that from 2006 till 2010 the process was smooth and without significant delays. The number of loans not funded started then to increase year after year until 2017, ending with approximately 15K loans to fund. This does not means delays got significantly worse year after year, but they're there.

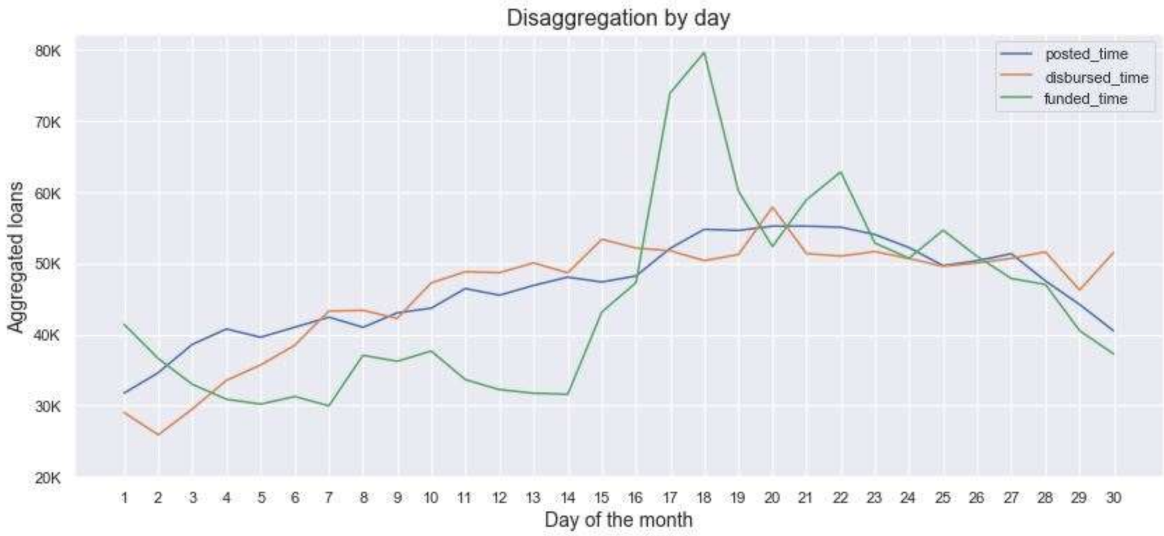
When looking at the disaggregation by month, we see the **cyclical nature** of the journey of Kiva loans. These cycles suggest that Kiva gather many loan requests before they post on their website. The fundings and disbursals follow the announcements (or, since the disbursal is the first to occur, the posting and funding follow the disbursals).



To further examine this cyclical nature, let us now display the aggregated values by month and finally, by day.



As we can see, there is a slight tendency to post, disburse and fund as the year comes to an end.



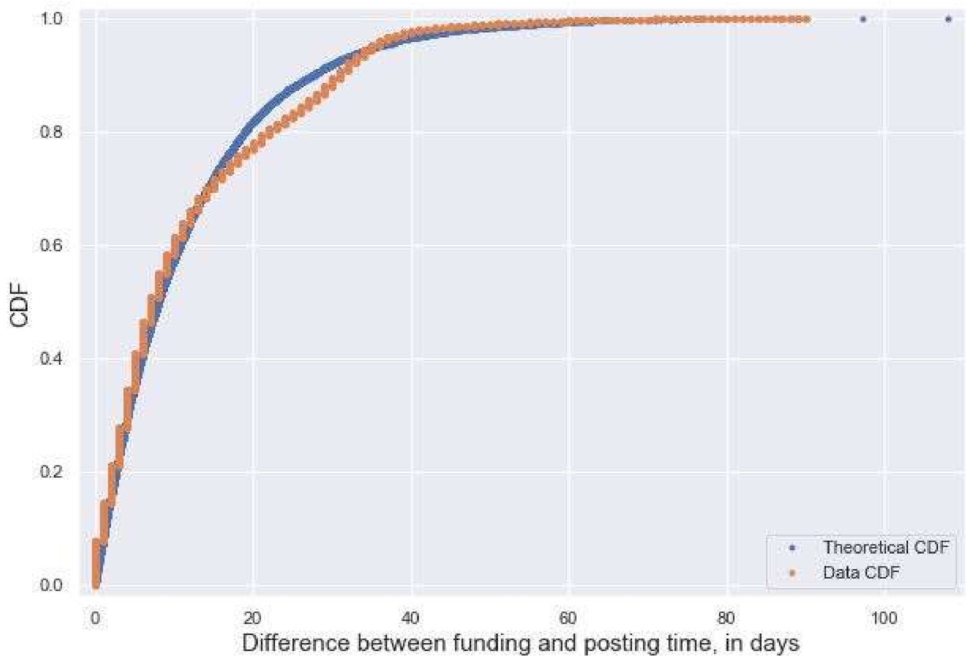
While the posting and disbursing periods relatively aggregate the same amount of loans throughout the month, there is a strong pressure to fund the loans during the **3rd week** of the month.

I then started to try to answer other questions:

I first focused on the time differences between the main variables:

98% of the differences between the time of disbursal and posting of loans occur on a range of -35 days and +32 days. Moreover, 90% of the differences are positive, that is, pre-disbursals. We could then immediately that at least 90% of the disbursals are made by field partners, the only ones who can pre-disburse.

When we look at the time a loan takes to get funded since the time it is posted on Kiva website, we see that 99% of them are funded within 50 days, and almost 90% within a month. In fact, this behavior follows an exponential distribution, as presented in the CDF:



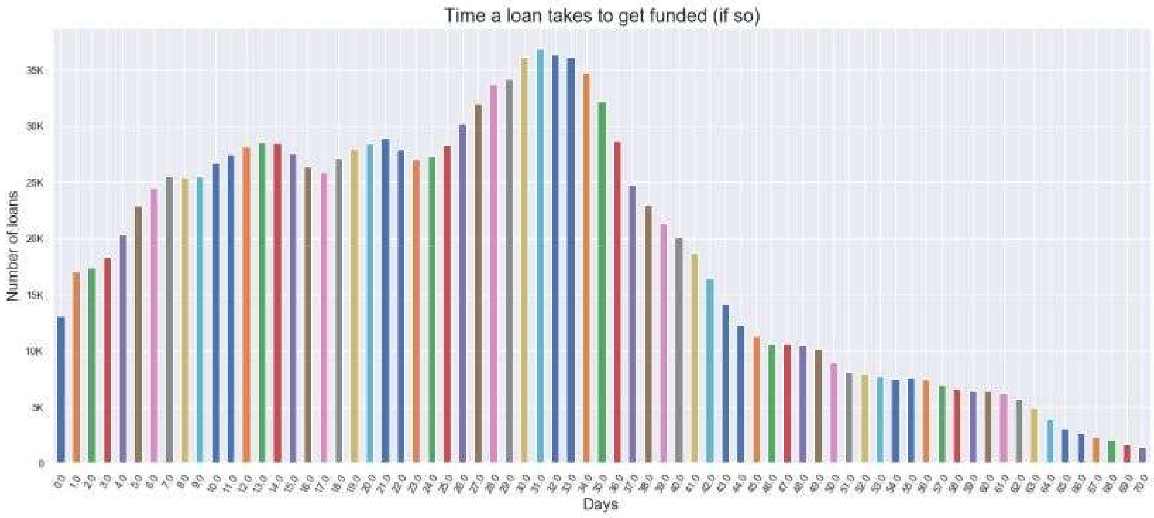
Close to 10% of the total loans are disbursed after the funding is complete, and 5% occur 14 days after the time of funding. This clear spike in the data suggests that 2 weeks could be a pre-staged arrangement. In most cases, a pre-disbursal occurs and only then the funding period commences. I also confirmed that the 5% cases where disbursal occurs 14 days after funding are distributed over time.

I then looked at the **real** time a loan takes to get funded, that is, looking at the difference between the funding time and the disbursal or posting time, depending on which of these occurred first. The following facts were checked:

- Close to 60% are funded within a month. Some irregularity (ups and downs) can be seen during this period.
- 97% are funded within 2 months.

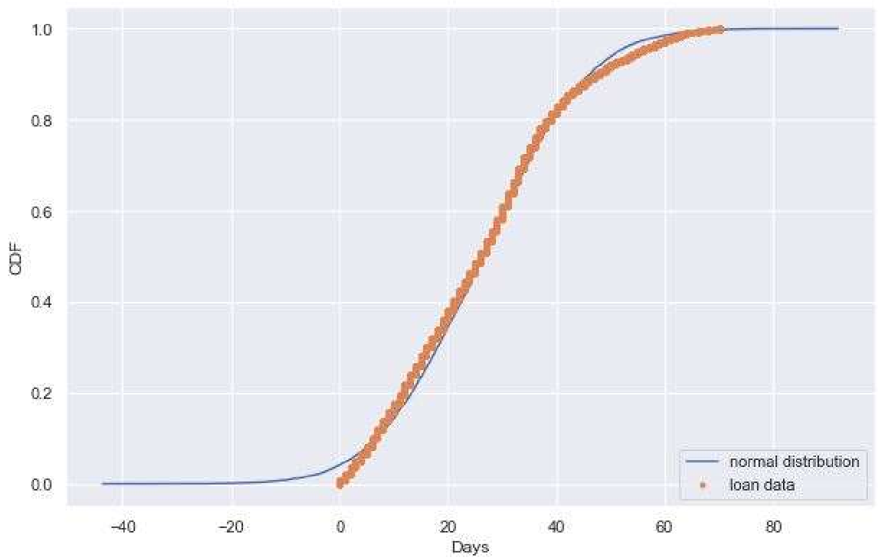
As a reminder, 8% of the total loans are only disbursed after they are funded (with 7% within two weeks). Considering these cases, this means that 99% of the loans are funded within 2 months and 2 weeks. Using D'Agostino's K-squared test, we reject the hypothesis that the funding period follows a normal distribution, with even a null p-value.

PDF:



However, this test encompasses negative values, so when looking at the ECDF and comparing it with the theoretical CDF of a normal distribution, the resemblance is more clear.

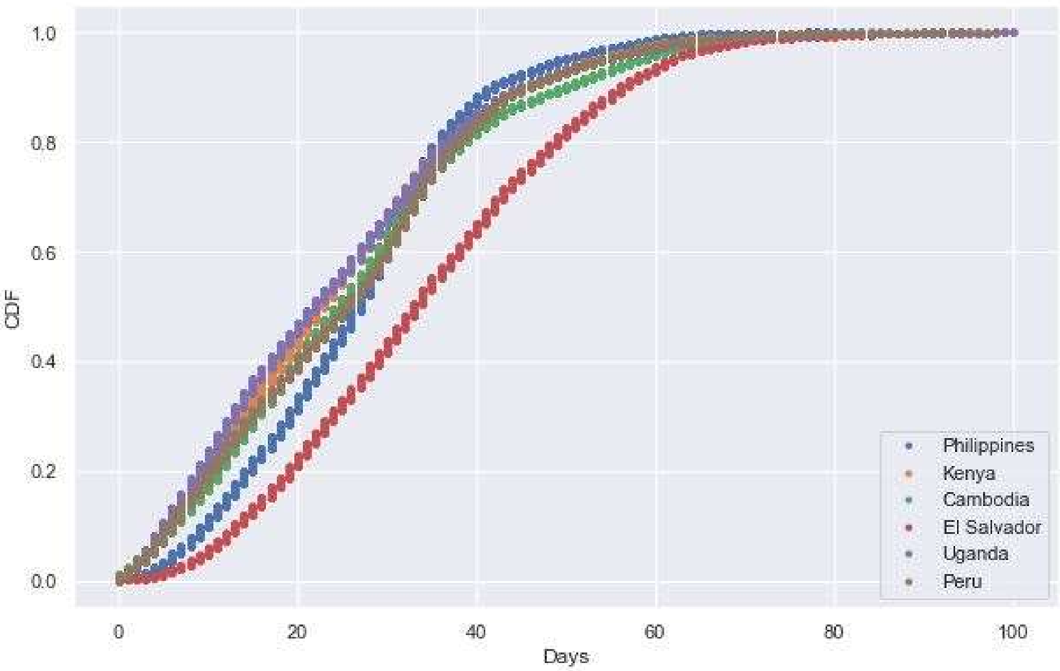
ECDF:



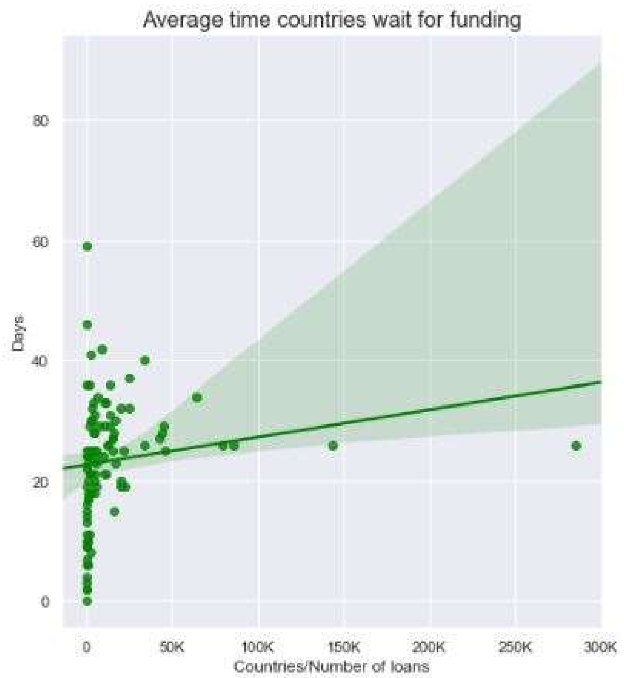
What may cause delays, then?

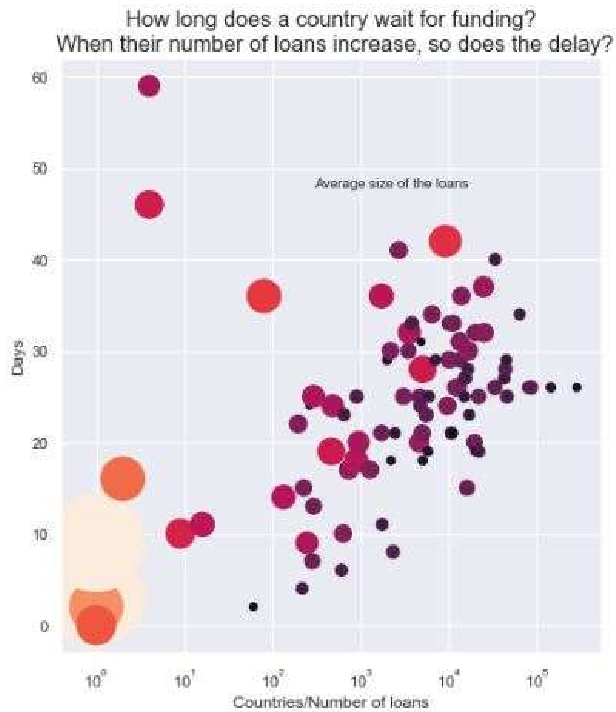
1. Do countries have influence regarding delays?

Half of the loans go to 6 countries (Philippines, Kenya, Peru, Cambodia, El Salvador, and Uganda) and 77% to 20. When looking specifically at the 6 countries PDFs and CDFs, we clearly distinguish El Salvador (takes 5% of the loans) from the other ones. It might be useful in further analysis. The Philippines (20%) could also be a motive for further analysis.



2. When a loan is funded (96% of the data), how long does it take to get that funding, on average, in each country? Does the number of loans have any impact?



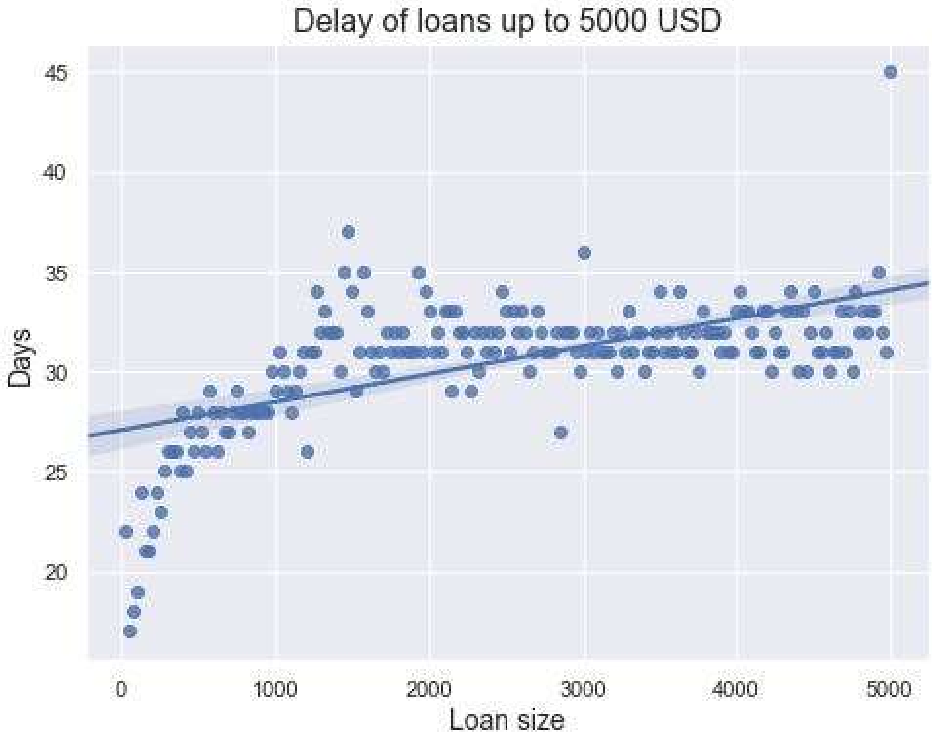


For countries with a few number of loans, the variation of average time to get the funding is significant. As soon as the number of loans starts to increase, the average number of days to get the funding rapidly get close to 30, where most of them reside. It seems to be that as the number of loans increases in a country, with all the remaining variables equal, so does the delay for funding a loan. There is a positive correlation between the two, but not significant when assuming a 5% significance level.

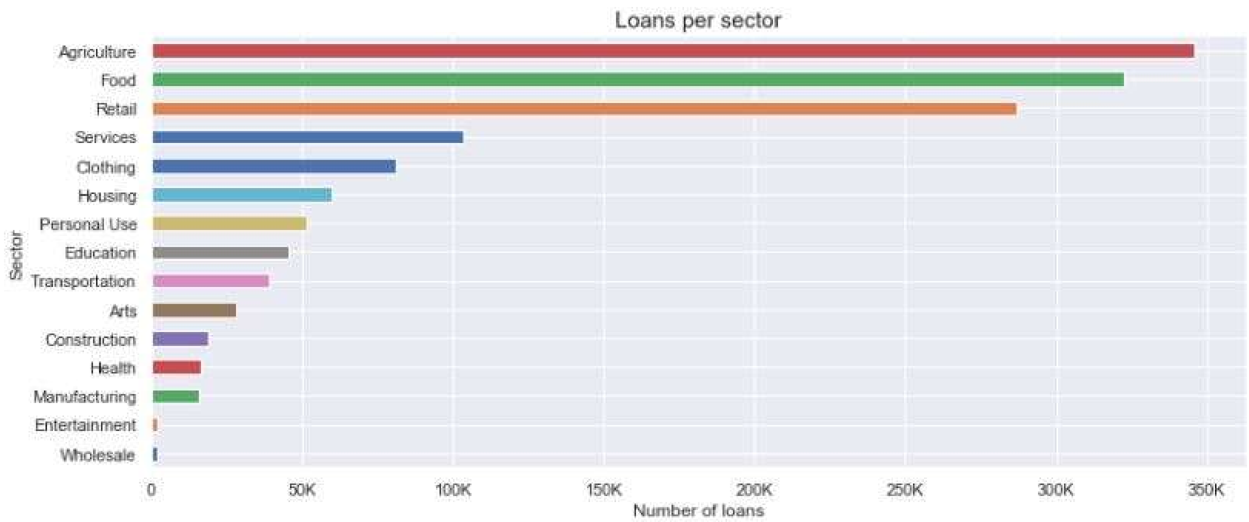
Given this result, however, other factors such as the characteristics of each country loan could be more influential, but we will keep this in mind. When filtering the countries that take a bit more than the average time to fund, USA, Colombia, Paraguay, Armenia, El Salvador, Lebanon, Bolivia, and Rwanda seem to have more impact.

3. When a loan is funded, do bigger loan amounts take more time to fund, on average?

Globally, there is a significant negative correlation (-16%) between the size of the loan and the time it takes to get funded, contrary to intuition. But 99% of the data has loans inferior to 5000 USD, and in that case, there is a positive correlation of 62%:

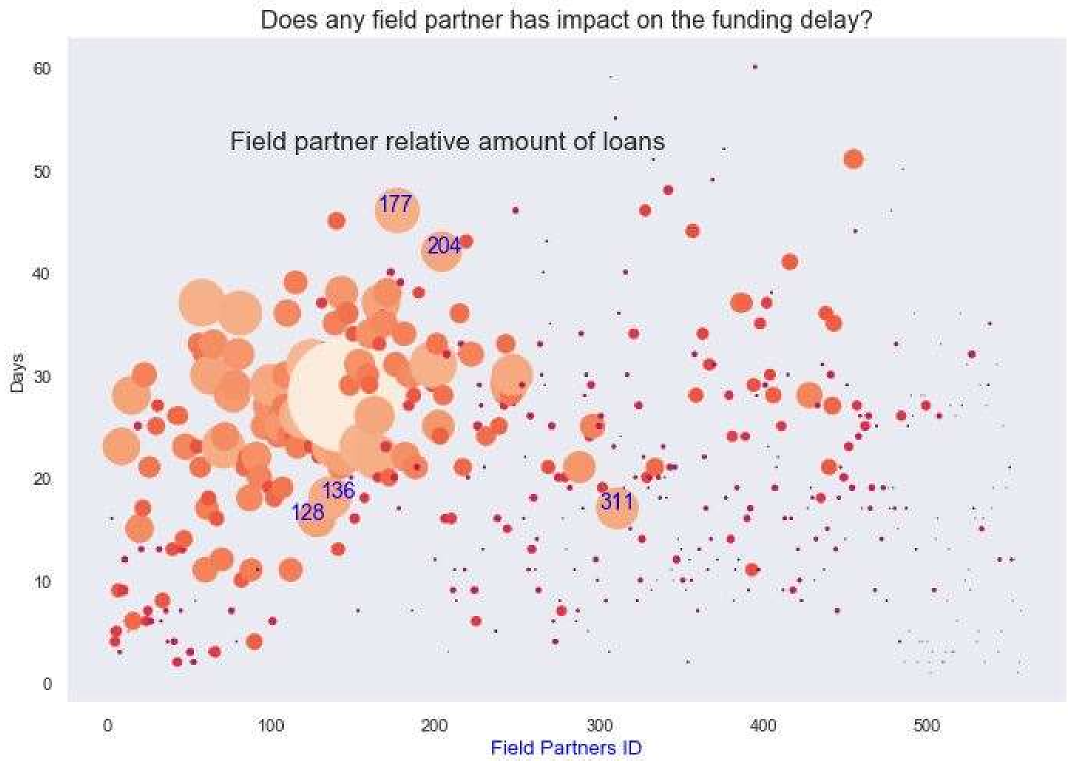


4. Does the sector of the loan influences the time of funding?



67% of the loans go to 3 sectors: Agriculture (24%), Food (23%) and Retail (20%). With further analysis regarding each sector main statistics, we can notice that the main 5 sectors (which gather 80% of the loans) are the ones who also influence the most a higher delay in fundings, where Agriculture has clearly the biggest impact.

5. How do Field partners fit in this analysis?

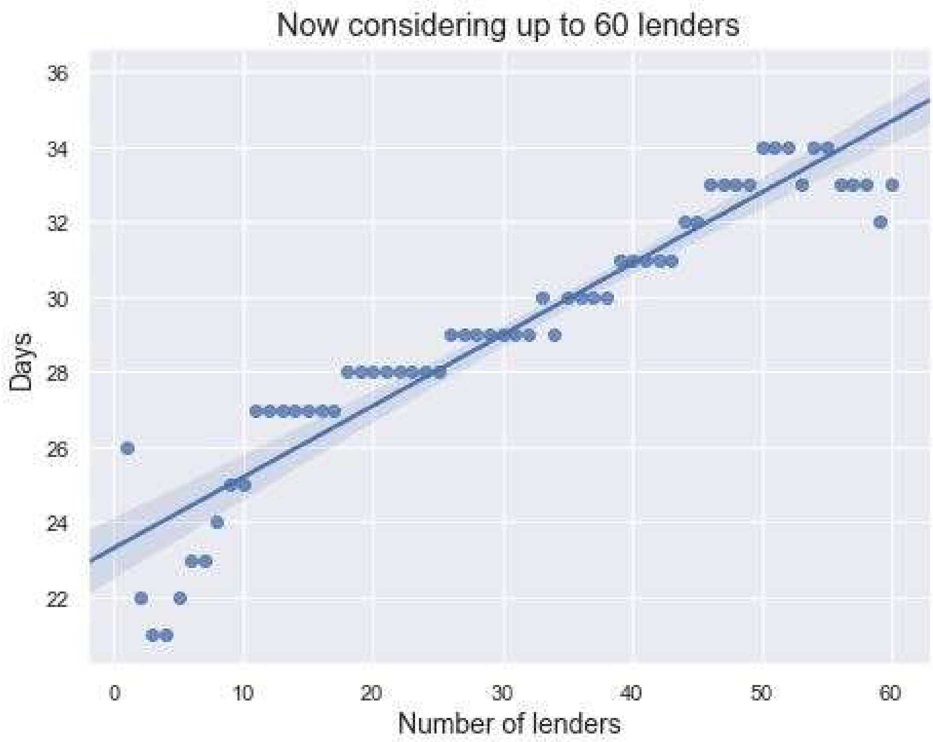


Of the 479 unique field partners, 11% of the loans are organized by field partner nº 145 and 90% by 120 field partners. There are not many field partners with a considerable size of loans that affect the overall average time loans take to get funded. Still, a few of them are relatively away from the concentrated 30-day delay. When filtering field partners that take more than 27 days on average to fund loans and with more than 1.5% of the loans, field partner with more loans, nº 145, is the one who causes more damage.

6. What is the relationship between the number of lenders and the time it takes to get funding?

Globally, as the number of lenders increased, the average number of days to fund a loan decreased.

However, 94% of the loans are composed by up to 60 lenders only. Here, there is a positive 95% correlation coefficient between the number of lenders and the time to fund a loan:

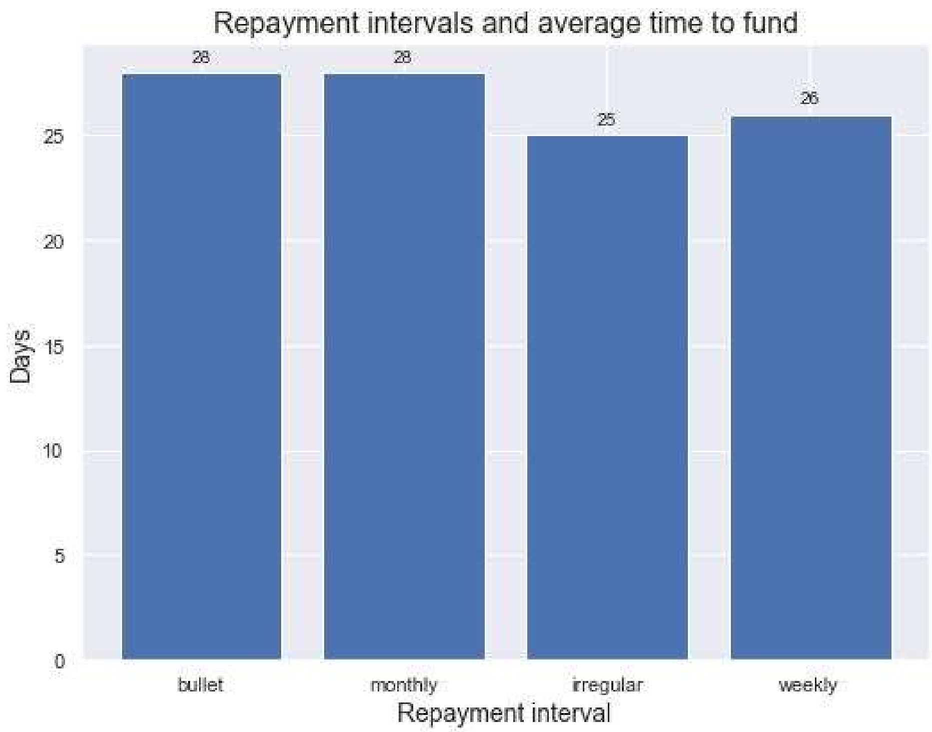


The data is presented in a seemingly organized disposition, which could suggest that the time to disburse and fund a loan is not subject to much uncertainty. In this sense, field partners may do prearranged commitments to fund the loans.

7. Does the repayment interval affect delay?

The main types of repayment intervals are “Monthly” (55%), “Irregular” (37%) and “Bullet”(8%):

There are slightly different statistics for each type. Dividing the data into two groups, one where the time to fund is below the median and the other above the median. Using a Chi-squared test, we conclude that there is a significant impact on the type of repayment interval. All in all, irregular repayments have a significantly better performance than monthly repayments, regarding delays.



8. When the loan is not completely funded, how big it is?

They were bigger than usual, as expected.