Capstone Project I

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 <).
  - It was also incoherent if the status appears as funded and there is a missing value in funded_time, but that was not the case.
  - 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 seem 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 would not be coherent if there was a missing 'disbursed_time' value and the variables 'funded_time' and status = funded were filled. Therefore, I **removed the cases when that was true** so that there is no ambiguity. The new number of missing values remained in the analysis.
- **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. If, however, there is a disbursal, the funded_amount column must be filled, but the opposite was not the case. 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.

2) Searching for outliers

2.1) '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**.

2.2) 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 used).

'funded_time' and 'disbursed_time'

I focused only on the cases where the disbursed_time was made after 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**.