Kiva.org assignment

Submitted by Team B

2 Aug 2019

# References

The Kiva Submission - Team B folder contains:

* Narrative
* R Scripts: Q1, Q3, Q5
* Pdf Exports of Final Plots: Rplot1, Rplot3a, Rplot3b, Rplot3c, Rplot3d, Rplot3e1, Rplot3e2, Rplot3f, Rplot3g, Rplot5

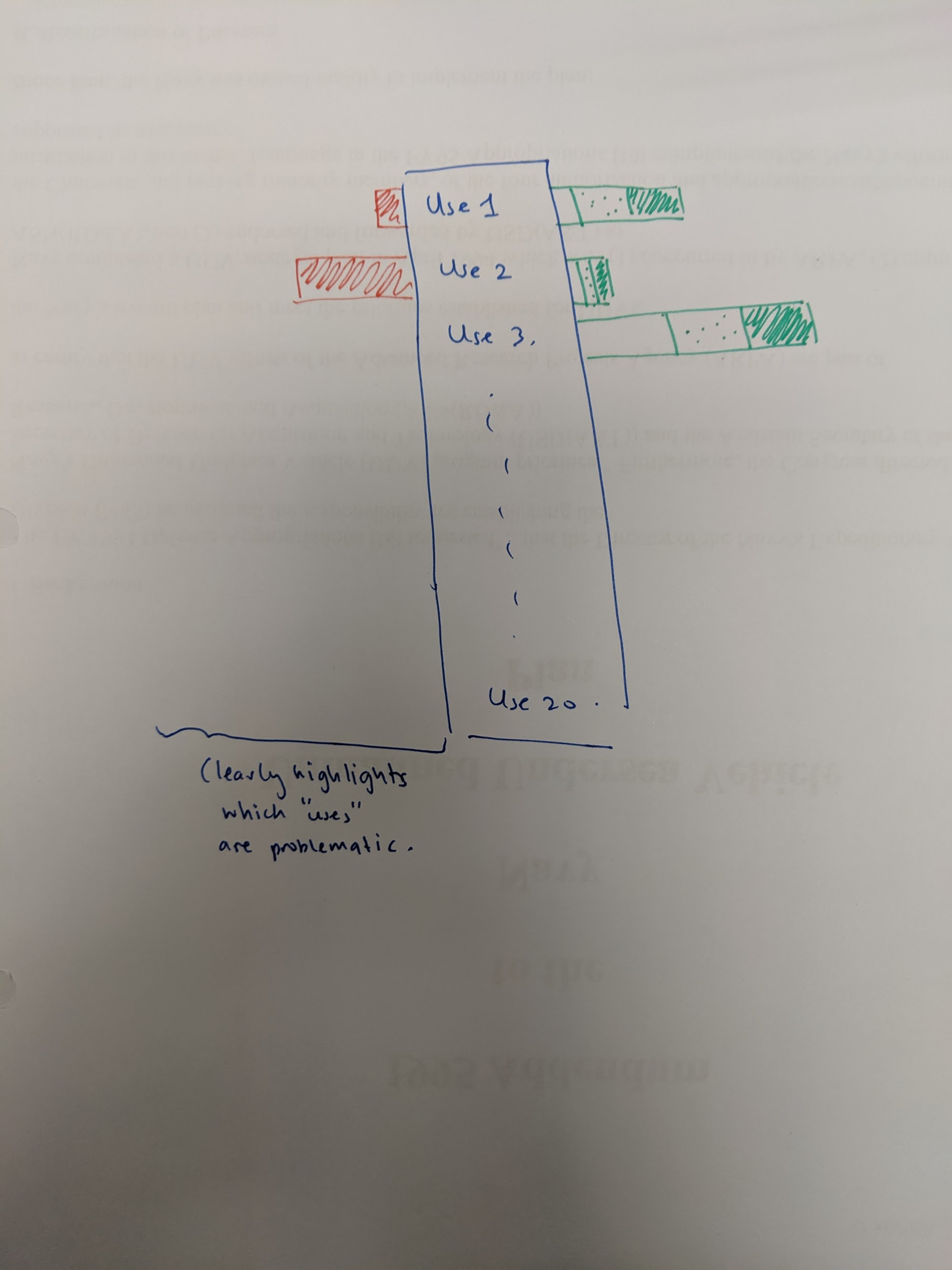
# Question 1

We correlated the loan taker’s “ability to pay” with their repayment\_intervals, with the “monthly”, “weekly” and “bullet” field options being generally good signs and the “irregular” field option signifying a lesser ability or inability to pay.

We intended for the code to

1. Rank the “Use” by number of occurrences (i.e. top 20 “Uses”).
2. Assign the following colour codes to the repayment\_interval field options:
   1. “bullet” (DARK GREEN)
   2. “weekly” (LIGHTER GREEN) (assuming ability to pay on a weekly basis means more disposable income cf. monthly),
   3. “monthly” (LIGHTEST GREEN),
   4. “irregular” (RED)
3. For each of the top 20 “Uses”, plot a stacked bar filled with the 4-colour gradient whose width corresponds to the sum of each field option - sort of like a Likert scale (there was a failed attempt to do sth with the likert package).

And for the final plot to look something like this



Actual plot differed greatly from expectations (See Rplot1.pdf). Some difficulties we ran into:

* Unexpectedly, it turned out that each “Use” has the same repayment\_interval. I.e. All 5000+ of the instances that the loans were used for “to buy a water filter...” are repaid on a monthly basis. We had mistakenly believed that there would be a mix of bullet, weekly, monthly and irregular repayment intervals.
* Inability to come up with the intended plot (e.g. the “Uses” were not properly ranked, colours are a mess, etc.) due to being R noobs and the lack of time to fumble around with R more.
* Unclean data - The “Uses” appear to have taken in data freely entered by the loan takers without any form of control or standardisation. There was therefore a great amount of duplication of the “Uses” in their various forms. A data cleansing step should have been added prior to the data manipulation. Unfortunately, please see 2nd bullet point again 😢

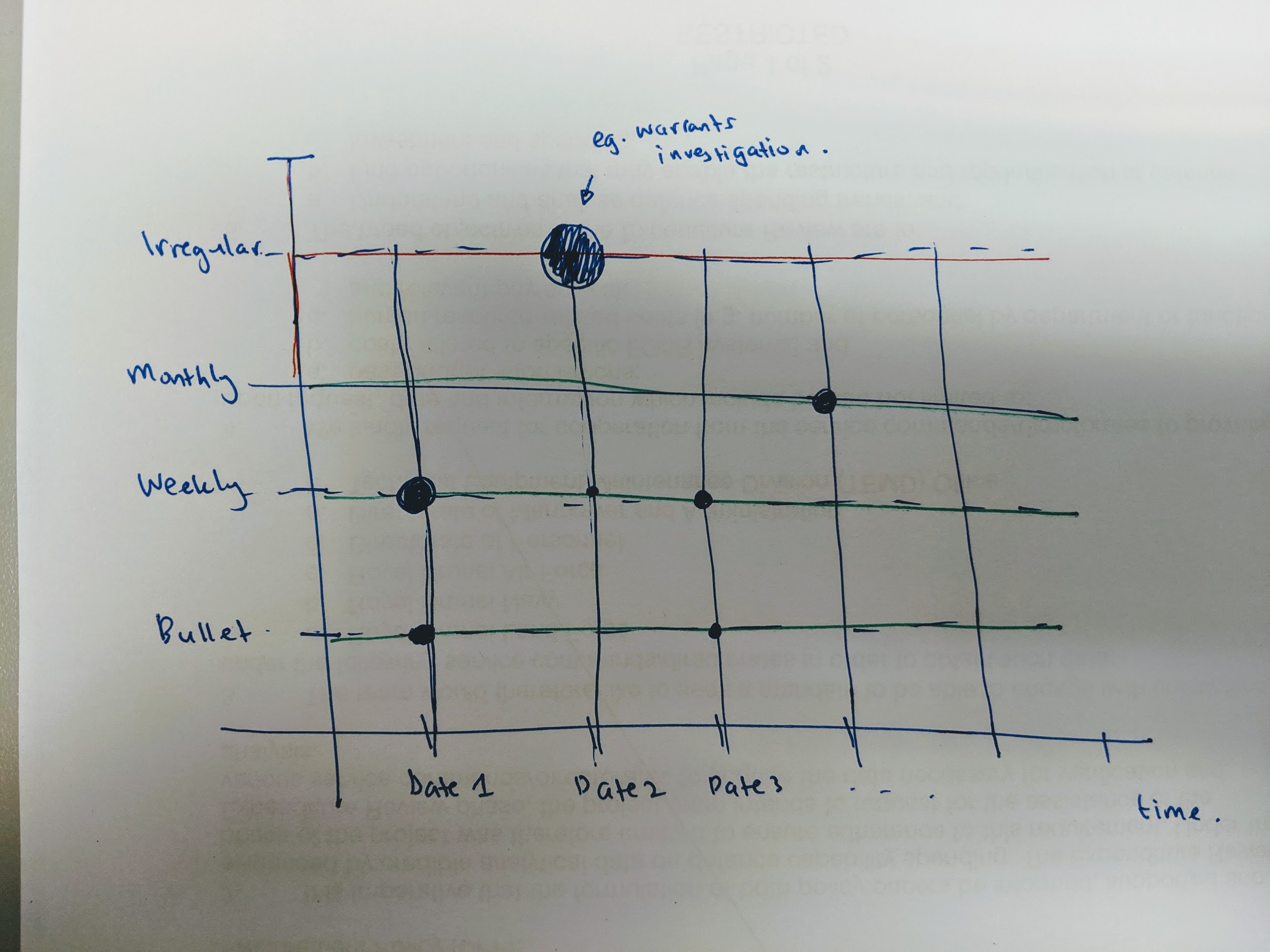
# Question 2

First, there was a need to define “high defaulting loans”. Default is the failure to meet the legal obligations or conditions of a loan. It is unclear whether Kiva imposes such obligations. An assumption we took to work within the available data was that “irregular” repayment\_intervals indicate “default”.

Coding intent

1. Plot a count plot of one of the various relevant kiva\_loan columns such as date against repayment\_interval.
2. x=time, y=repayment interval
3. Observe if there are any trends surrounding points that fall within the “irregular” section repayments and time e.g. coinciding with certain events/times of the year (e.g. payday, recession).
4. Observe and investigate any outliers.

How we pictured the final plot



# Question 3

In an attempt to observe interesting statistics segregated by the regions, below are some of the relationships assessed:

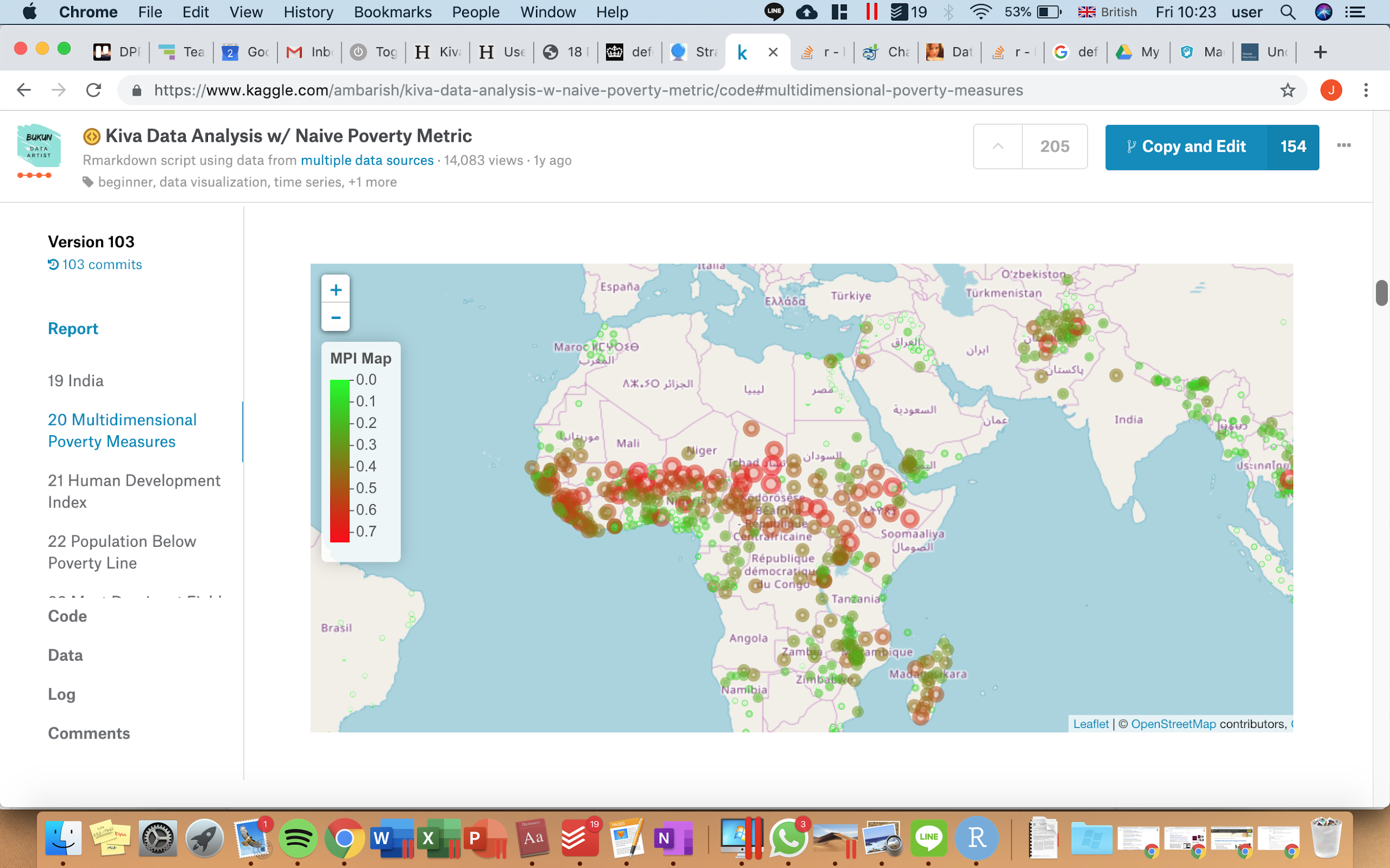
1. To observe the different loan sector over the different world region;
2. To observe the different loan sector over the continents;
3. To observe the number of loans taken over the different regions of different countries;
4. To observe the different loan type over the region of the different countries;
   1. Loan type needs to be further clean.
5. To observe the different loan sector dispersion in certain countries;
   1. This may be visually better presented if overlay with the country’s map
6. To observe the amount of loan for different sectors in different countries; and,
7. To observe the amount of loan for different sectors in different regions.
   1. This will need the loan amount to be normalised.

# Question 4

Coding intent

1. Use Kaggle Dataset “Multidimensional Poverty Measures”, which investigates poverty in the world. According to Kiva, multidimensional poverty measures reveal who is poor and how they are poor – the range of different disadvantages they experience.\*\*Higher the MPI, poorer is the country\*\*
2. Overlay the data from kiva\_loans with the Kaggle data on multidimensional poverty measures.
3. Use a heatmap of the world where the conditions are as follows
   1. “Greener” colours would be a Low MPI + High funded\_amount
   2. “Amber” colours would be High MPI + Low funded\_amount; Low MPI + Low funded\_amount
   3. “Redder” colours - High MPI (Very Poor Country) + High funded\_amount

How we pictured the final plot



(Something like this, but with different (extra) conditions for the colour maps ofc)

# Question 5

Part 1: Analysing gender discrepancy in the number of loans being funded

The code analysed the gender groups for each loan that was being funded, which are divided into the groups below:

1. Individual female group
2. Individual male group
3. All female group
4. All male group
5. Mixed group (Both female and male)

As a result, a bar chart was plotted, comparing the number of loans that are being funded for each group. Individual female group are funded with the loans much more than the other 4 groups, as seen in the bar chart plotted.

The code also analysed the total number of females vs males overall, which resulted in another bar chart being plotted. Overall, females are funded with loans more than males.

Part 2: Analysing gender discrepancy in the amount of loans being funded

Intended to analyse the amount of loans being funded but could not figure out an efficient code to convert funded\_amount (which is in different currencies) to a common currency.