Project proposal

Daniel and Filo inc.

1. Introduction

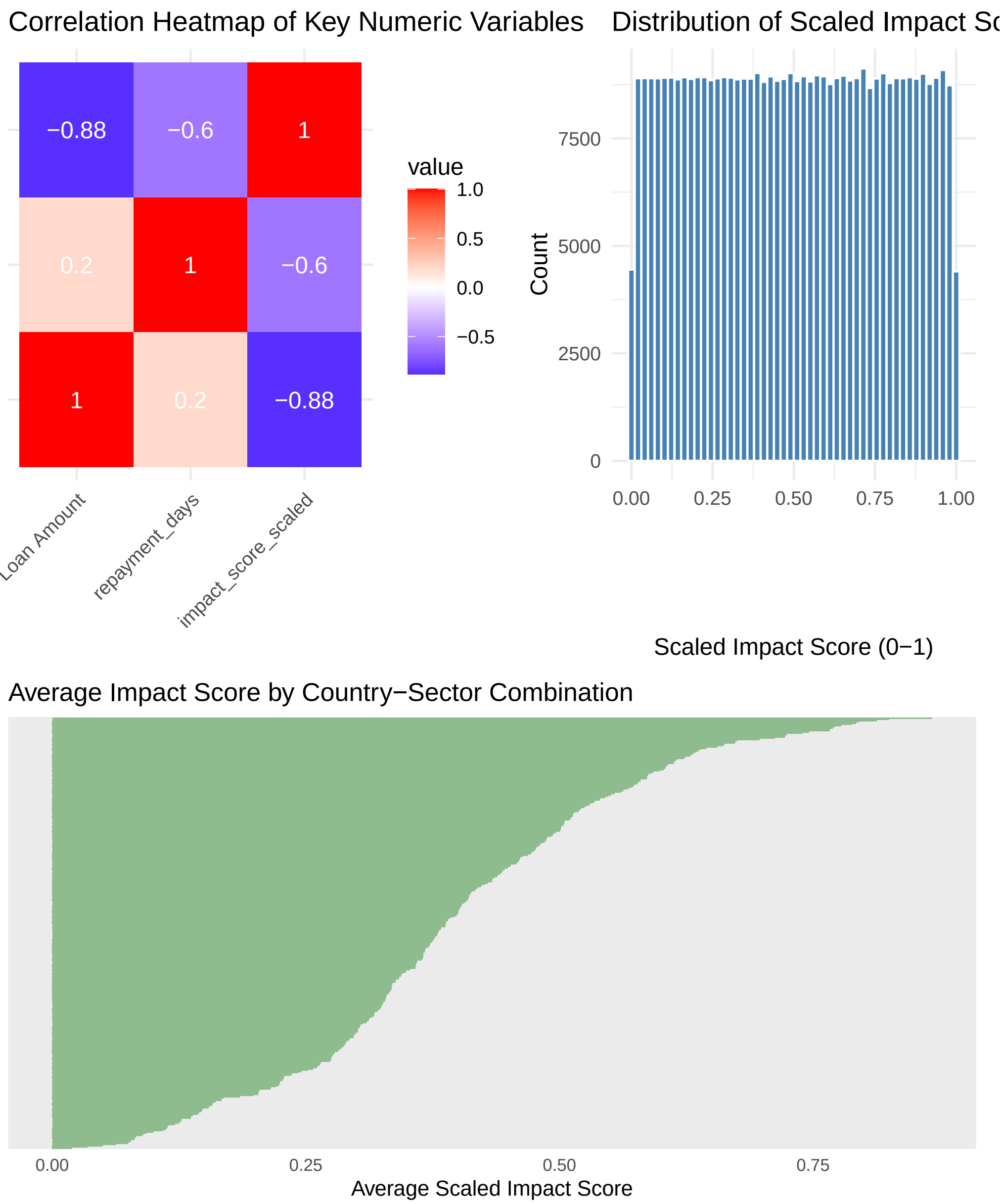
Kiva is a global nonprofit organization that facilitates microlending by connecting individual lenders with low-income borrowers across developing regions. Since its founding in 2005, Kiva has helped finance millions of loans to support entrepreneurship, education, agriculture, and more — often in communities underserved by traditional financial institutions. The platform's high repayment rates and transparency have made it a popular tool for individuals seeking to generate positive social impact through small-scale lending. Yet, while the platform is widely used, many individual lenders still face a central challenge: how to allocate limited funds where they will do the most good. For these lenders, it is not just about supporting a borrower — it's about supporting the right loan: one that returns quickly, reliably, and with minimal investment. This project seeks to address this challenge by our Research Question: Which combinations of loan sector and borrower country are associated with the highest repayment impact — defined as fully repaid, small, and fast-returned loans? Loans that are repaid quickly are considered more impactful from our perspective because they demonstrate that the borrower was able to generate sufficient income from the investment in a short timeframe, allowing the capital to be recycled and re-lent to other borrowers more efficiently. By calculating this score across hundreds of thousands of historical loans, we aim to uncover actionable insights into the profiles of high-impact loans. This analysis moves beyond default risk prediction and offers a new lens for guiding efficient and impactful microlending decisions on the Kiva platform.

2. Data

The dataset used in this project originates from Kiva.org, a global crowdfunding platform that provides microloans to underserved borrowers across the world. The full dataset was compiled and shared by Dr. Moran Koren for academic research purposes and consists of several tables containing detailed loan-level, payment, schedule, and country-level metadata. For the purposes of our analysis, we focused on the main big_table.csv, which aggregates information about each loan, including borrower country, sector, loan amount, loan status, and dates related to funding and repayment. We further processed this dataset to calculate repayment indicators and construct an "impact score" that reflects how efficient and successful each loan was in terms of amount, duration, and repayment status. After filtering and transformation, we retained only the most relevant fields for answering our research question. All transformations are documented in the code and reflected in the final working dataset. All variables were selected based on their relevance to the research question and were processed to ensure completeness and consistency. The structured dataset was saved as a .csv file in the /data folder and is accompanied by a README.md file describing each column.

3. Preliminary results

In our initial exploration, we examined the structure and distribution of the key variables used to evaluate loan impact. Specifically, we focused on how impact scores relate to loan amount, repayment time, and borrower characteristics such as country and sector. We visualized the correlation between numeric variables (loan amount, repayment duration, and impact score), and confirmed a strong negative relationship between repayment time and impact, as expected. To better understand what combinations tend to produce more impactful loans, we calculated the average impact score for each country-sector combination (e.g., “Philippines - Retail”, “Kenya - Agriculture”). The resulting barplot highlights all such combinations with more than 100 loans, and demonstrates substantial variation in repayment impact across regions and industries. For readability, axis labels were removed from the chart. The histogram Distribution of scaled impact score shows the distribution of scaled impact scores (0–1), indicating a roughly uniform spread across the normalized range. This confirms that the rank-based scaling successfully differentiates loans in terms of impact, enabling fair comparison across loan types.



4. Data analysis plan

Outcome (Y):

Our outcome variable is the scaled impact score (impact_score_scaled), a continuous metric that reflects the effectiveness of a loan based on its repayment completeness, duration, and size. This score is normalized between 0 and 1, allowing fair comparisons across all loans.

Predictors (X):

We use two categorical explanatory variables:

- `Country - the borrower's country`
- `Sector - the economic sector associated with the loan`

We are particularly interested in the interaction between these two factors to identify high-performing combinations.

Comparison Groups:

Each unique Country–Sector combination forms a comparison group. We will examine the average impact score within each group to determine which combinations yield the most effective loans in terms of return efficiency.

Methods:

We will begin with descriptive visualizations to explore the distribution of impact scores across sectors and countries. Next, we will conduct a Two-Way ANOVA, using Country and Sector as factors, to assess:

- `The main effect of Country`
- `The main effect of Sector`
- `The interaction effect between Country and Sector`

If the ANOVA reveals statistically significant differences, we will apply Tukey's HSD post-hoc test to determine which specific country–sector pairs differ meaningfully from others.

Expected Results:

We expect that both Country and Sector will significantly influence the impact score, and that some specific combinations (e.g., "Philippines – Retail") will outperform others

Teamwork

Since we are working as a pair, we decided to collaborate on both the conceptual and technical aspects of the project. While we will each take responsibility for different subtasks, we intend to complete the statistical analysis and interpretation together to ensure a shared understanding and cohesive results.

Appendix

README – Kiva Loan Dataset

Each row in the dataset represents a single loan made through the Kiva platform. The dataset aggregates and joins information from multiple original tables, including loan-level details, country metadata, and payment records.

Variables

- **loan_amount**: *(numeric)* Total amount of money requested by the borrower (in USD).
- **repayment_gap**: *(numeric)* Difference between the total scheduled repayment and the actual total paid. Positive values indicate underpayment.
- **sector**: *(categorical)* Broad economic category of the loan (e.g., Agriculture, Retail, Education).
- **region**: *(categorical)* Geographic region of the borrower (e.g., Sub-Saharan Africa, South Asia).
- **repaid_fully**: *(binary)* Equal to 1 if the loan was fully repaid or overpaid, 0 otherwise.

Source code

```r
library(knitr)
library(tidyverse)
library(broom)
library(htmltools)
library(patchwork)
# Load libraries
library(tidyverse)
library(data.table)
library(lubridate)

# Load only the main file
loans <- fread("/Users/danielkravtsov/Downloads/kivadata_org/big_table.csv")

# Parse dates
loans <- loans %>%
  mutate(
    funded_date = as.Date(`Funded Date`),
    paid_date = as.Date(`Paid Date`)
  )

# Create repayment indicator and duration
loan_data_impact <- loans %>%
  filter(!is.na(funded_date), !is.na(paid_date), !is.na(`Loan Amount`)) %>%
  mutate(
    repaid_fully = ifelse(Status == "paid", 1, 0),
    repayment_days = as.numeric(difftime(paid_date, funded_date, units = "days")),
    impact_score = ifelse(repaid_fully == 1 & repayment_days > 0,
                          1 / (`Loan Amount` * repayment_days),
                          0)


  ) %>%
  select(
    id,
    Country,
    Sector,
    `Loan Amount`,
    funded_date,
    paid_date,
    Status,
    repaid_fully,
    repayment_days,
    impact_score
  )

loan_data_impact <- loan_data_impact %>%
  mutate(
    impact_score_scaled = case_when(
      impact_score == 0 ~ 0,
      impact_score > 0 ~ percent_rank(impact_score)
    )
  )



library(reshape2)

# 1. Correlation heatmap for key numeric variables
numeric_vars <- loan_data_impact %>%
  select(`Loan Amount`, repayment_days, impact_score_scaled)

cor_matrix <- cor(numeric_vars, use = "complete.obs", method = "spearman")
heatmap_data <- melt(cor_matrix)

p1 <- ggplot(heatmap_data, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  geom_text(aes(label = round(value, 2)), color = "white", size = 4) +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +
  labs(title = "Correlation Heatmap of Key Numeric Variables", x = "", y = "") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))


# 2. Histogram of impact score (scaled using percent_rank)
p2 <- ggplot(loan_data_impact, aes(x = impact_score_scaled)) +
  geom_histogram(fill = "steelblue", color = "white", bins = 50) +
  labs(
    title = "Distribution of Scaled Impact Score",
    x = "Scaled Impact Score (0-1)",
    y = "Count"
  ) +
  theme_minimal()


# 3. Combine Country and Sector into one variable
loan_data_impact <- loan_data_impact %>%
  mutate(country_sector = paste(Country, Sector, sep = " - "))

# 4. Calculate average impact score for each Country-Sector combo (with >100 loans)
p3 <- loan_data_impact %>%
  group_by(Country, Sector) %>%
  filter(n() > 100) %>%
  summarise(mean_impact = mean(impact_score_scaled, na.rm = TRUE)) %>%
  mutate(combo = paste(Country, Sector, sep = " - ")) %>%
  ggplot(aes(x = reorder(combo, mean_impact), y = mean_impact)) +
  geom_col(fill = "darkseagreen") +
  coord_flip() +
  labs(
    title = "Average Impact Score by Country-Sector Combination",
    x = "",  # axis text removed for clarity
    y = "Average Scaled Impact Score"
  ) +
  theme_minimal(base_size = 10) +
  theme(axis.text.y = element_blank())  # hide Y-axis labels for readability
```

```
(p1 | p2) / p3
```