Module 1 – R Practice Assignment

Ayush Anand

Northeastern University: College of Professional Studies ALY 6010: Probability Theory and Introductory Statistics

Professor XYZ

September 27th, 2024

About the Dataset: Loan Transactions

- Overview: The dataset comprises a comprehensive record of all private loans issued under USAID's Development Credit Authority (DCA) since its inception in 1999.

- Data Protection: To safeguard the privacy of borrowers and bank partners, all sensitive and personally identifiable information has been removed from the dataset.

- Loan Characteristics: The dataset includes various attributes such as loan amounts, disbursement dates, business sectors, and indicators for woman-owned businesses and first-time borrowers.

- Analysis Potential: Exploratory data analysis (EDA) on this dataset can reveal trends in loan distribution, identify the impact of being woman-owned or a first-time borrower on loan amounts, and provide insights into sector-specific financing patterns.

- Significance: This data serves as a crucial resource for understanding how loans are allocated and their relationship with different business characteristics, contributing to informed decision-making in financial planning and policy development.

Data Processing Steps

1. Package Installation and Library Loading:

   - Installed necessary packages: `stats`, `dplyr`, `skimr`, and `DataExplorer`.

   - Loaded these libraries into the R environment.

```
install.packages("stats")
install.packages("dplyr")
# Load necessary libraries
```

2. Data Import:

   - Read the dataset from a CSV file using `read.csv()` and stored it in the variable `data`.

```
data <- read.csv("C:/Users/ayush/OneDrive/Documents/R_Assignment_1_Stats/RStudio/data_1.csv")
```

3. Initial Data Exploration:

   - Assigned the imported data to `df1` for further processing.

   - Used `str()` to check the structure of the data and `head()` to view the first few rows.

```r
df1 <- data
# Check the structure of the data
str(data)
```

4. Date Conversion:

- Converted the `Disbursement.Date` and `End.Date` columns to Date format using `as.Date()` for accurate date analysis.

```r
# Convert Disbursement Date and End Date to Date format
df1$Disbursement.Date <- as.Date(df1$Disbursement.Date, format="%m/%d/%Y")
df1$End.Date <- as.Date(data$End.Date, format="%m/%d/%Y")
```

5.Handling Missing Values:

- Replaced missing values in the `Business.Sector` column with "Unknown" to avoid loss of data.

- Verified the changes using `summary()`.

```r
# Verify that missing values have been replaced
summary(df1$Business.Sector)
```

6. Statistical Summary:

- Obtained summary statistics for the `Amount..USD.` column to understand its distribution.

- Created scaled versions of the loan amount (in thousands and millions) for easier interpretation.

```r
# Check summary statistics for Amount (USD)
summary(df1$Amount..USD.)

# If scaling is required (i.e., dividing by 1000 or 1,000,000 for better interpretability):
df1$Amount.USD.K <- df1$Amount..USD. / 1000  # Amount in thousands
df1$Amount.USD.M <- df1$Amount..USD. / 1000000  # Amount in millions

# Check the summary of scaled data
summary(df1$Amount.USD.K)
summary(df1$Amount.USD.M)
```

7. Outlier Treatment:

- Checked and removed outliers in `Is.Woman.Owned` and `Is.First.Time.Borrower` columns by setting invalid entries to NA.

```r
# Remove outliers in Is Woman Owned?
df1$Is.Woman.Owned <- ifelse(df1$Is.Woman.Owned %in% c(0, 1), df1$Is.Woman.Owned, NA)

# Remove outliers in Is First Time Borrower?
df1$Is.First.Time.Borrower <- ifelse(df1$Is.First.Time.Borrower %in% c(0, 1), df1$Is.First.Time.Borrower, NA)
```

8. Frequency Distribution:

- Generated frequency distributions for woman-owned businesses and first-time borrowers using `table()`.

```
#Step 1: Frequency Distribution of Woman-Owned Businesses and First-Time Borrowers
# Frequency distribution for Is Woman Owned?
table(df1$Is.Woman.Owned)

# Frequency distribution for Is First Time Borrower?
table(df1$Is.First.Time.Borrower)
```

9. Cross-Tabulation Analysis:

- Created cross-tabulations to explore relationships between `Business.Sector` and ownership/borrower status.

- Aggregated loan amounts based on `Is.Woman.Owned` and `Is.First.Time.Borrower`.

```
# Cross-tabulation between Business Sector and Is Woman Owned
woman_owned_vs_sector <- xtabs(~ Business.Sector + Is.Woman.Owned, data=df1)
print(woman_owned_vs_sector)

# Cross-tabulation between Business Sector and Is First Time Borrower
first_time_vs_sector <- xtabs(~ Business.Sector + Is.First.Time.Borrower, data=df1)
print(first_time_vs_sector)

# Cross-tabulation between Is Woman Owned and Loan Amount
woman_owned_vs_loan <- aggregate(Amount..USD. ~ Is.Woman.Owned, df1, mean)
print(woman_owned_vs_loan)

# Cross-tabulation between Is First Time Borrower and Loan Amount
first_time_vs_loan <- aggregate(Amount..USD. ~ Is.First.Time.Borrower, df1, mean)
print(first_time_vs_loan)
```

10. Trend Analysis:

- Analyzed trends in loan amounts over time by disbursement and end dates using `aggregate()`.

- Visualized these trends with line plots using `ggplot()`.

11. Data Visualization:

- Created various plots, including:

  - A histogram to show the distribution of loan amounts.

  - Bar plots to visualize the distribution of loans by business sector and the comparison of woman-owned vs. non-woman-owned businesses.

  - Pie charts to illustrate the percentage of woman-owned vs. non-woman-owned businesses and first-time vs. repeat borrowers.

12. Loan Duration Calculation

- Calculated loan duration in days by subtracting `Disbursement.Date` from `End.Date`.

13. Correlation Analysis:

- Assessed correlations between loan amounts and borrower characteristics using `cor()`.

Frequency distribution

Frequency distribution for Is Woman Owned?

table(df1$Is.Woman.Owned)

```
> #Step 1: Frequency Distribution of Woman-Owned Businesses and First-Time Borrowers
> # Frequency distribution for Is Woman Owned?
> table(df1$Is.Woman.Owned)

     0      1
149254  37263
>
```

Frequency distribution for Is First Time Borrower?

table(df1$Is.First.Time.Borrower)

```
> # Frequency distribution for Is First Time Borrower?
> table(df1$Is.First.Time.Borrower)

     0      1
125451  61067
>
```

Cross-Tabulation

Cross-tabulation between Business Sector and Is Woman Owned

woman_owned_vs_sector <- xtabs(~ Business.Sector + Is.Woman.Owned, data=df1)

print(woman_owned_vs_sector)

```
> #Step 2: Cross-Tabulation of Woman-Owned Businesses and First-Time Borrowers with Loan Amount and Business Sector
> # Cross-tabulation between Business Sector and Is Woman Owned
> woman_owned_vs_sector <- xtabs(~ Business.Sector + Is.Woman.Owned, data=df1)
> print(woman_owned_vs_sector)
                                    Is.Woman.Owned
Business.Sector                          0     1
                                      7791   162
  Agriculture                        59433 17318
  Construction                         366    46
  Education                            549   389
  Energy                               255   174
  Fisheries/Aquaculture                863   296
  Forestry/Wood                        156    31
  Health                               708   314
  Housing                               71    19
  Information & Communication Technologies   129    22
  Infrastructure                       110     2
  Manufacturing                      16789  1398
  Other Service                       6101  5243
  Tourism                              176   146
  Trade/Commerce                     55125 11511
  Transportation                       632   192
> |
```

Cross-tabulation between Business Sector and Is First Time Borrower

first_time_vs_sector <- xtabs(~ Business.Sector + Is.First.Time.Borrower, data=df1)

print(first_time_vs_sector)

```
> # Cross-tabulation between Business Sector and Is First Time Borrower
> first_time_vs_sector <- xtabs(~ Business.Sector + Is.First.Time.Borrower, data=df1)
> print(first_time_vs_sector)
                                    Is.First.Time.Borrower
Business.Sector                          0     1
                                      7884    70
  Agriculture                        37710 39041
  Construction                         254   158
  Education                            445   493
  Energy                                77   352
  Fisheries/Aquaculture                270   889
  Forestry/Wood                        124    63
  Health                               611   411
  Housing                               42    48
  Information & Communication Technologies    74    77
  Infrastructure                        70    42
  Manufacturing                      16899  1288
  Other Service                       3426  7918
  Tourism                              123   199
  Trade/Commerce                     56895  9741
  Transportation                       547   277
> |
```

Cross-tabulation between Is Woman Owned and Loan Amount

woman_owned_vs_loan <- aggregate(Amount..USD. ~ Is.Woman.Owned, df1, mean)

print(woman_owned_vs_loan)

```
> # Cross-tabulation between Is Woman Owned and Loan Amount
> woman_owned_vs_loan <- aggregate(Amount..USD. ~ Is.Woman.Owned, df1, mean)
> print(woman_owned_vs_loan)
  Is.Woman.Owned Amount..USD.
1              0    13692.113
2              1     4059.032
> |
```

Cross-tabulation between Is First Time Borrower and Loan Amount

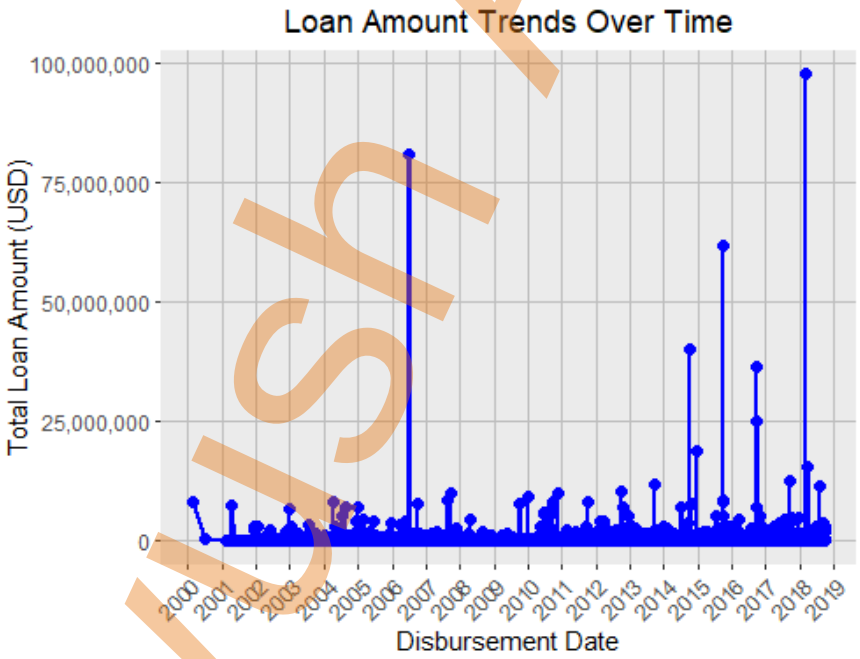
first_time_vs_loan <- aggregate(Amount..USD. ~ Is.First.Time.Borrower, df1, mean)

print(first_time_vs_loan)

```
> # Cross-tabulation between Is First Time Borrower and Loan Amount
> first_time_vs_loan <- aggregate(Amount..USD. ~ Is.First.Time.Borrower, df1, mean)
> print(first_time_vs_loan)
  Is.First.Time.Borrower Amount..USD.
1                      0     12580.59
2                      1     10230.25
> |
```

Visualization

Analyzing trends in loan amounts by disbursement date



Key Insights:

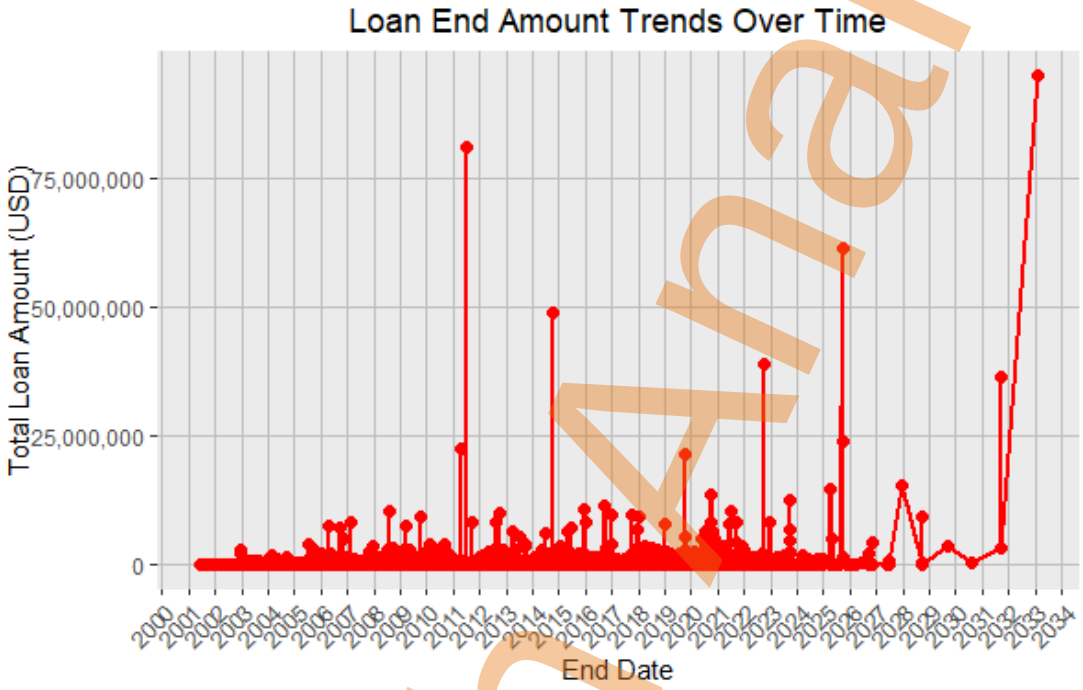
Significant Fluctuations: The plot exhibits substantial variability in loan amounts over time, suggesting periods of high and low lending activity. This could be influenced by various factors such as economic conditions, interest rates, or policy changes.

Outlier Analysis: There appear to be a few outliers with exceptionally high loan amounts. These could represent large-scale projects, government initiatives, or other unusual

circumstances. Further investigation into these outliers might provide valuable insights into the factors driving such large-scale lending.

Potential Clustering: While a more detailed analysis would be required, there might be evidence of clustering or periodicity in the data. For instance, there could be cyclical patterns related to economic cycles or seasonal factors. Exploring these potential patterns could help identify recurring trends in loan disbursement.

Analyzing trends in loan amounts by end date



Key insights:

High Variability in Loan Amounts Over Time:

There are significant fluctuations in the total loan amounts at different end dates. Some dates see very large spikes (notably above 75 million USD), indicating that certain periods had considerably higher loan disbursements compared to others.

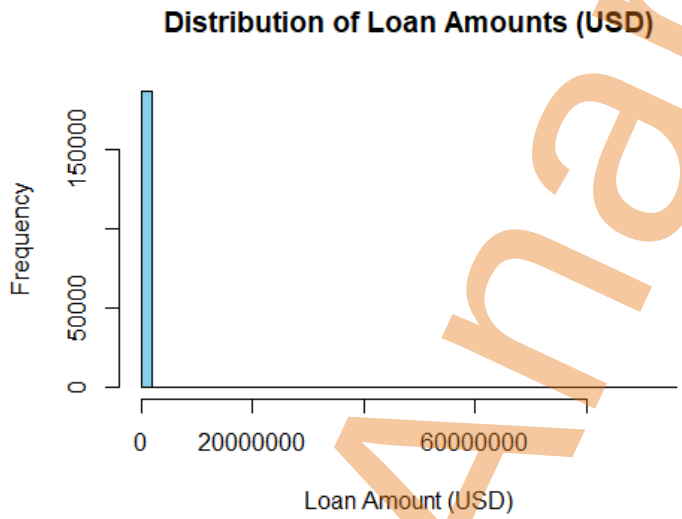
Recent Surge in Loan End Amounts:

Towards the far right of the plot, there is a noticeable and steep upward trend in loan amounts, with a large spike near the latest end dates. This suggests a dramatic increase in loan closures in the most recent period.

Multiple Isolated Loan Spikes:

Several isolated spikes in the total loan amount are visible throughout the timeline. These peaks likely represent significant individual loans or clusters of loans that were finalized on those specific end dates, indicating particular periods of high-value loan disbursements.

Histogram for Loan Amount (USD)



Key Insights:

Extremely Skewed Distribution:

The plot indicates a highly skewed distribution of loan amounts, with most loans concentrated at very low values, close to zero. This suggests that the majority of the loans in the dataset are relatively small in size.

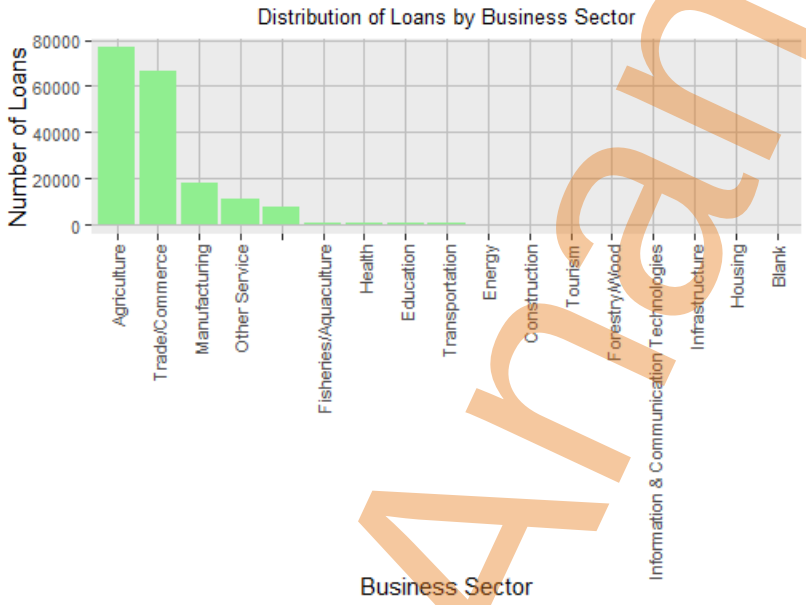
Outliers in Loan Amounts:

Although the bulk of the data is concentrated near smaller loan amounts, there are some outliers visible far to the right of the x-axis (in the millions). These outliers represent a few significantly large loans compared to the majority.

High Frequency of Small Loans:

The frequency of loans in the lower loan amount range is incredibly high, peaking around the smallest loan amounts. This may indicate that small-scale lending is far more common in the dataset compared to large loans.

Distribution of Loans by Business Sector



Key Insights:

Dominance of Agriculture:

The Agriculture sector stands out significantly, with the highest number of loans, surpassing 80,000. This indicates that the majority of loans in the dataset are granted to businesses in the agricultural sector.

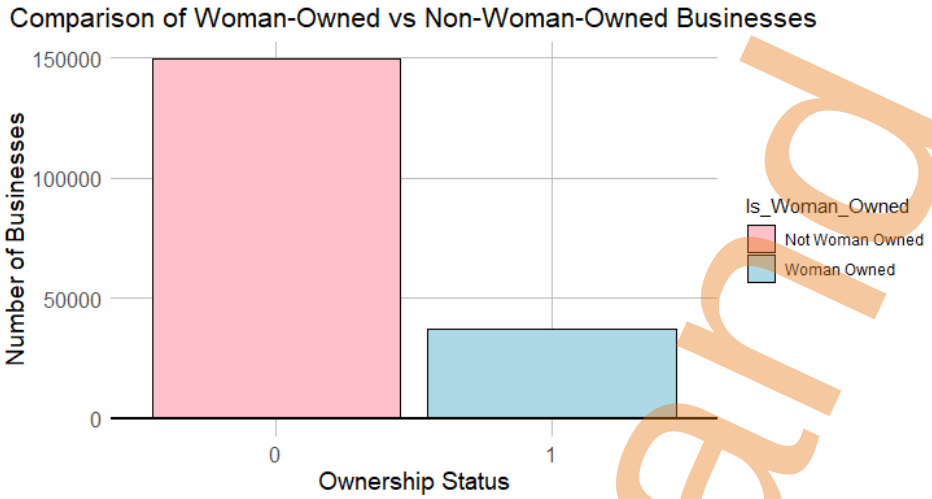
High Loan Count in Trade/Commerce and Manufacturing:

After Agriculture, the Trade/Commerce and Manufacturing sectors have the second and third highest number of loans, though they trail behind Agriculture. This shows that these sectors also receive a substantial amount of financial support but not to the same extent as Agriculture.

Sparse Loans in Other Sectors:

The plot demonstrates that sectors such as Health, Energy, Education, and Information & Communication Technologies receive comparatively fewer loans, with some sectors having very little to no loan representation. This could suggest a lower demand for or allocation of loans in these areas.

Bar chart comparing and Count the number of woman-owned vs non-woman-owned businesses chart comparing woman-owned vs non-woman-owned businesses

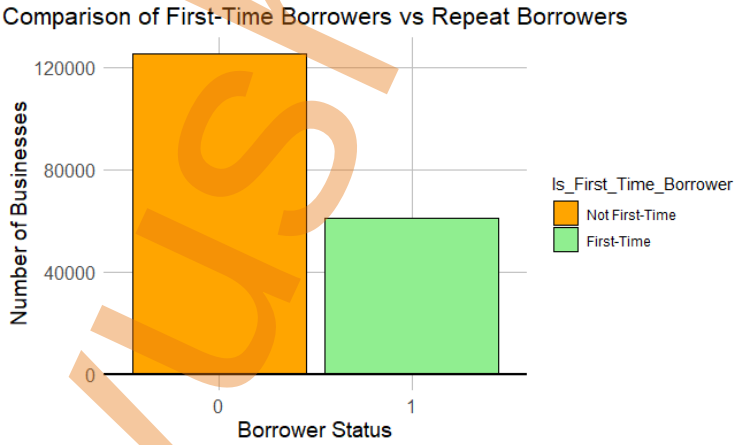


Key Insights:

Prevalence of Non-Woman-Owned Businesses: The plot indicates a significant disparity between the number of non-woman-owned and woman-owned businesses, suggesting that barriers or challenges may hinder women's entrepreneurship.

Gender Gap in Business Ownership: The comparison highlights a clear gender gap in business ownership, with non-woman-owned businesses outnumbering woman-owned businesses.

Bar chart comparing and count the number of first-time borrowers vs repeat borrowers

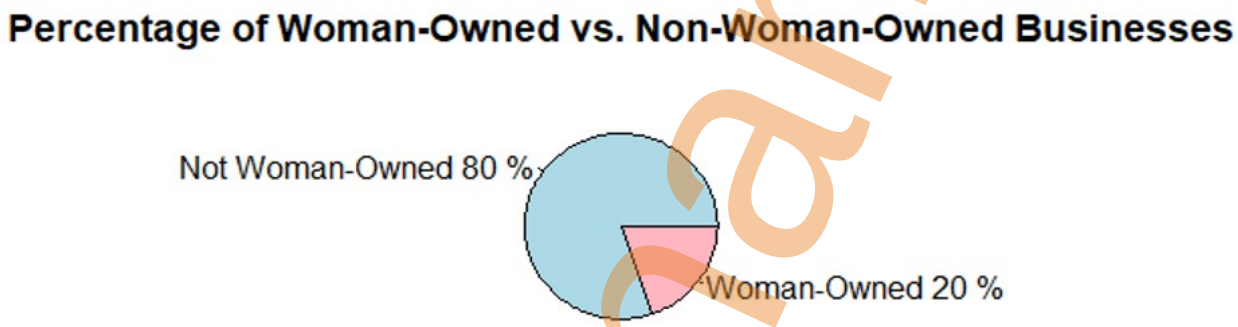


Key Insights:

- **Prevalence of Repeat Borrowers:** The chart indicates that a significantly higher number of businesses are repeat borrowers compared to first-time borrowers. This suggests that a substantial portion of businesses have established relationships with lenders and continue to seek financing.

- **Repeat Borrowers Outnumber First-Time Borrowers:** The comparison highlights a clear disparity between the two groups, with repeat borrowers far exceeding first-time borrowers. This could be attributed to various factors, such as the success of previous loans, trust in lenders, or the availability of financing options for established businesses.

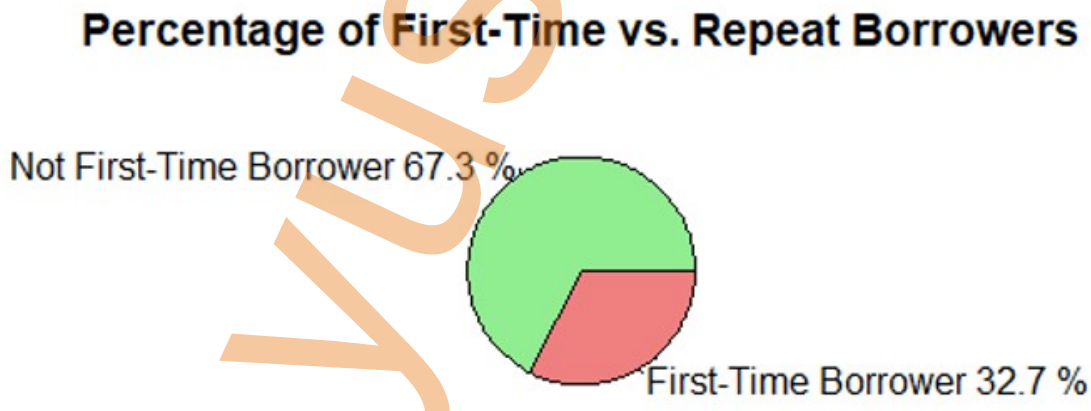
Pie chart of loan amount for woman-owned businesses



Key Insights:

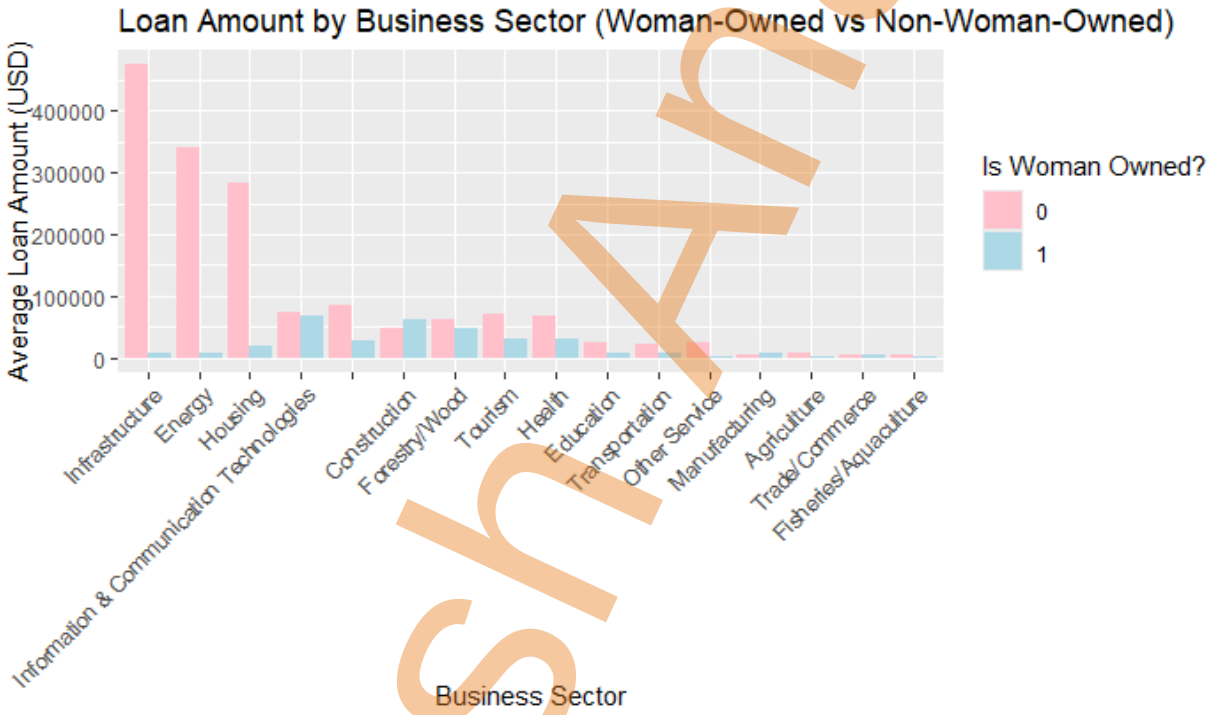
- **Significant Disparity:** The chart clearly demonstrates a significant gender gap in business ownership. Non-woman-owned businesses constitute a much larger proportion (80%) compared to woman-owned businesses (20%).
- **Dominance of Non-Woman-Owned Businesses:** The large size of the pie chart segment representing non-woman-owned businesses visually emphasizes their dominance in the dataset. This suggests that there may be systemic barriers or challenges that hinder women's entrepreneurship.

Pie chart of loan amount for first-time borrowers



- **Key Insights:**
  Prevalence of Repeat Borrowers: The chart indicates that a significantly higher percentage of borrowers are repeat borrowers (67.3%) compared to first-time borrowers (32.7%). This suggests that a substantial portion of businesses have established relationships with lenders and continue to seek financing.
- **Dominance of Repeat Borrowers:** The larger size of the pie chart segment representing repeat borrowers visually emphasizes their dominance in the dataset. This could be attributed to various factors, such as the success of previous loans, trust in lenders, or the availability of financing options for established businesses

Bar plot comparing average loan amount by sector for woman-owned vs non-woman-owned businesses



Key Insights:

- **Disparities in Loan Amounts:** The chart reveals significant differences in average loan amounts between woman-owned and non-woman-owned businesses across various sectors. This suggests that there may be systemic barriers or biases that hinder women's access to larger loans.
- **Sector-Specific Differences:** The comparison highlights variations in loan amounts between different business sectors, regardless of ownership status. This indicates that the industry in which a business operates can play a significant role in determining the level of financing it receives.

Summary

The dataset includes private loan records from USAID's Development Credit Authority (DCA) since 1999, with sensitive information removed to safeguard borrowers. Loan amounts, company sectors, disbursement dates, and indicators for woman-owned firms and first-time borrowers are all important aspects to consider. The study focuses on loan distribution trends, ownership patterns, and industry-specific lending information.

Key Findings:

Loan Distribution Patterns: Agriculture receives the most loans, followed by trade/commerce and manufacturing, while sectors such as health and education receive fewer.
Loan amount trends: Significant swings in loan amounts over time, with discrete spikes indicating large-scale lending. The distribution is heavily weighted toward smaller loans.
Ownership Insights Women-owned enterprises and first-time borrowers are underrepresented relative to their male counterparts, indicating potential access hurdles.
This dataset gives useful insights into loan allocation and aids in financial and policy decision-making.

Reference

Data is taken from DATA.GOV

https://catalog.data.gov/dataset/development-credit-authority-dca-data-set-loan-transactions-a8dbe