Exploring Financial Inclusion Disparities Across the World

Report Prepared By:

Sri Krishna Yerramilli
Jahnavi Ravi
Rishabh Kansal
Yeshwanth Pendyala

Submitted to:

Dr. Dragos Bozdog
School of Business

FA 582: Foundations of Financial Data Science

Project Report

Spring 2023

Table of Contents

Abstract

This report summarizes the modeling and analysis results associated with the Financial Inclusion data science project. The goal of the project is to analyze global financial inclusion data and identify the key socioeconomic factors behind these disparities. It also aims to analyze US data to use as a benchmark. To answer these questions, we compared data from 140 countries across 5 primary indicators including account ownership rates and debit/credit card usage rates during 4 different years over the last decade. In 2021, 76% of adults had an account at a bank or a regulated institution and global account ownership increased by 50% from 2011 to 2021. We also used unsupervised learning methods like KNN and Decision Trees to classify countries based on their economies and corresponding financial inclusion data. The overall accuracy of Decision Tree model is better than that of KNN, indicating that the data might be better suited for classification compared to clustering. Overall, extensive data analysis was used to analyze and identify several recurring trends since the machine learning models did not result in very high accuracy.

Keywords

Financial Inclusion; Exploratory Data Analysis; Clustering; Tree-structured Classifier; Unsupervised Learning

1 Introduction

Financial inclusion is a critical aspect of economic development and societal progress, aiming to provide individuals and businesses with access to affordable and reliable financial services. Understanding the dynamics of financial inclusion rates and the factors that influence them is crucial for policymakers, financial institutions, and organizations working towards achieving inclusive economic growth. In this context, data analysis plays a pivotal role in uncovering patterns, trends, and insights that can inform decision-making and shape strategies for promoting financial inclusion.

The primary dataset used in this project is the Global Findex Database, a comprehensive collection of financial inclusion indicators published by the World Bank. This dataset encompasses a wide range of information, including account ownership, card usage, savings, internet access, phone ownership, and financial resilience, covering over 1,000 indicators across countries worldwide. By leveraging this dataset, we aim to delve into the factors that contribute to or hinder financial inclusion, identify trends and correlations, and gain a deeper understanding of the global landscape of financial access. To ensure the quality and relevance of the analysis, a meticulous data collection process was undertaken. Preprocessing of the data involved categorizing indicators, addressing missing data, and standardizing the format to enable consistent analysis.

In this study, we employ a combination of exploratory data analysis, correlation analysis, and summary statistics to gain insights into the relationship between various indicators of financial inclusion. The exploratory data analysis focuses on a subset of countries carefully selected to represent diverse financial inclusion rates and socioeconomic factors. By analyzing the relationships between different indicators, we can identify patterns and correlations that may indicate underlying drivers or barriers to financial inclusion.

Furthermore, we delve into the variations in financial inclusion rates based on gross national income (GNI) per capita and continents. By grouping the indicators according to these factors, we can discern differences across income groups and geographical regions. We also compare the financial inclusion statistics in the United States with the averages of different income group economies in the world to get a better understanding of the disparities in financial inclusion. Finally, machine learning models like KNN and Decision Tree were used to try and correctly classify countries based on different financial inclusion indicators These insights enable us to identify areas where interventions and policy measures are needed to bridge gaps in financial inclusion and promote equitable access to financial services.

Overall, this study harnesses the power of data analysis to explore and illuminate the complex landscape of financial inclusion, with the aim of informing evidence-based policies and strategies that can drive meaningful change and pave the way for a more inclusive and prosperous future.

2 Data Collection

2.1 Data Source

The primary dataset used in this project is the Global Findex Database published by the World Bank. The dataset provides information for over 1000 indicators on topics such as account ownership, card usage, saving, internet access, phone ownership, and financial resilience. A specific dataset containing country-level data in an Excel (.xls) file was downloaded from https://www.worldbank.org/en/publication/globalfindex/Data. The country-level dataset was chosen for its extensive nature and to enable comparison across different countries around the globe. The file also contained a series table providing a detailed description of the different indicators and their aggregation method. Subsequently, the dataset was loaded as a data frame for preprocessing using the "readxl" package in R.

2.2 Preprocessing

The data was first checked manually. Although they were over 1000 indicators, a quick look through the data was enough to classify them into 10-15 broad categories. Furthermore, we were able identify some indicators that were more crucial for analysis using a mixture of basic understanding of the project's goals and common sense. Additionally, the dataset contained a lot of missing data. Since this data could not be imputed, we were able to further narrow our list of indicators based on which indicators the most data available while discarding the rest.

We converted all the numbers in the dataset into a numeric type and into percentages values to maintain a standardized format. We also converted columns into factors with different levels when it was appropriate to do so. For example, the year column was converted into a factor of 4 levels: 2011, 2014, 2017 and 2021. Since we wanted to investigate and compare financial inclusion statistics across different countries, we added a "continent" column to the original dataset. This was done using the "countrycode" package in R, as seen in Fig.1, that uses the "iso3c" 3-letter country codes to classify countries into 5 continents: Africa, Americas, Asia, Europe, and Oceania.

```
# Add a new column called "Continent"
data_ind$Continent <- suppressWarnings(countrycode(sourcevar = data_ind$Code, origin = "iso3c", destination =
"continent", nomatch = NA))

unmatched_code <- which(is.na(data_ind$Continent))
data_ind <- data_ind[-unmatched_code,]
```

Fig. 1: *Code to add "continent" column to dataset.*

For our exploratory data analysis and summary statistics, we only wanted to consider the most relevant data. For this purpose, we only included data from 2021. We decided to include 26 countries that were carefully chosen to provide a diversified representation of the varying financial inclusion rates as well as other socioeconomic factors. Figure 2 below provides a sample of the dataset used for exploratory data analysis. Although this dataset only contained 26 rows and 14 columns, extensive

analysis and machine learning clustering and classification was carried out utilizing the data from all 140 countries across 4 years, making it a much bigger dataset.

	Country	Code	Year	Income group	Account	Borrowed Money	Fin Inst Account	Used card	Owns card	Inactive account	Internet	Saved at Fin Inst or Money account	Owns Phone	Saved money
1	Albania	ALB	2021	Upper middle income	0.4417417	0.4268601	0.4417417	0.08430479	0.2709842	0.0489169173	0.7813150	0.09658711	0.9060962	0.3178514
2	Argentina	ARG	2021	Upper middle income	0.7162709	0.5156083	0.6632544	0.43522075	0.5647530	0.0349251218	0.8516717	0.14346325	0.9207171	0.3871555
3	Australia	AUS	2021	High income	0.9932315	0.6700200	0.9932315	0.94388384	0.9713351	0.0006703335	0.9378852	0.69173324	0.9545861	0.8298064
4	Brazil	BRA	2021	Upper middle income	0.8403575	0.5875807	0.8356319	0.55264652	0.7012470	0.0444609262	0.7829549	0.25374699	0.8466303	0.4624065
5	Canada	CAN	2021	High income	0.9963462	0.8612603	0.9963462	0.96713209	0.9792484	0.0043441867	0.9441248	0.63909453	0.8823817	0.7814819

Fig. 2: *Dataset used for Exploratory Data Analysis*

3 Exploratory Data Analysis

3.1 Indicators

Fig. 3 below contains a list of the 10 indicators we used to generate our correlation matrix and also produce some basic statistical summaries. The detailed explanation for each indicator is provided next to it in the table. All of the data collected by the Findex Database that is used in this project is surveyed from individuals above the age of 15.

account_t_d	Account (% age 15+)
borrow_any	Borrowed any money (% age 15+)
fin1_t_d	Financial institution account (% age 15+)
fin2_7_t_d	Owns a debit or credit card (% age 15+)
fin4_8_t	Used a debit or credit card (% age 15+)
fin9N_10N_t_d	Has an inactive account (% age 15+)
Internet	Has access to the internet (% age 15+)
fin17a_17a1_d	Saved at financial institution or using a mobile money account (% age 15+)
Own_phone	Owns a mobile phone (% age 15+)
save_any	Saved any money (% age 15+)

Fig. 3: *List of 10 Indicators used for correlation matrix and summary statistics*

3.2 Correlation Matrix

We generated a correlation matrix (Fig. 4) using the 10 indicators listed above in order to get a better understanding of the relationship between different pairs of indicators. This type of analysis helped us identify the indicators with the strongest correlation to each other, which we then used for further analysis, to generate summary statistics and eventually build our machine learning models.

As expected, having an account is negatively correlated with all of the other indicators. However, it seems to be negatively correlated the most to indicators such as having internet access and using a credit or debit card. This makes sense intuitively as well because limited internet access makes it difficult to access a bank account and an inactive account generally means that person is not using their credit or debit cards.

Apart from this, there seems to be a strong positive correlation between 4 indicators: owning an account, owning an account at a financial institution, owning a credit/debit card, and using a credit/debit card. Moving forward, these are the primary indicators we use for our analysis.

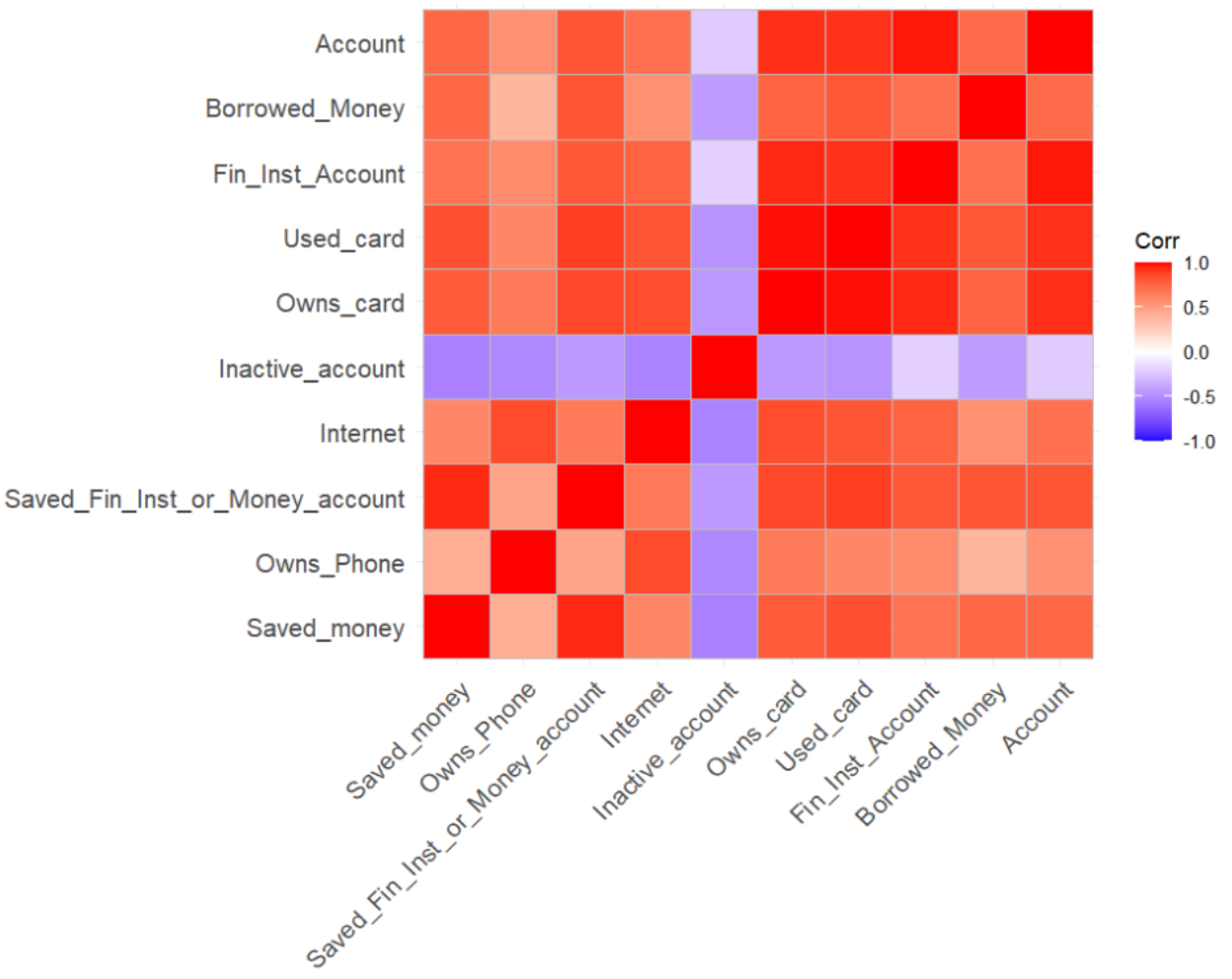

Fig. 4: *Correlation Matrix*

4 Basic Summary Statistics

4.1 Overview

For all of our groupwise summary statistics, we applied the "summaryBy" function in R and used a helper function that we defined outside in order to calculate different statistics like the number of data points, minimum, maximum, mean and median. Fig. 5 and Fig. 6 contains snippets of the code for both these functions along with our implementation.

```
summary_acc <- summaryBy(Account ~ Income_group, final_data, FUN = siterange)
colnames(summary_acc) <- c("Income","Count","Min","Max","Mean","Median")
```

Fig. 5: *"summaryBy" function from "doBy" library in R*

```
# Function for group wise statistics
siterange <- function(x){
  c(length(x),min(x),max(x),mean(x),median(x))
}
```

Fig. 6: *Helper function to calculate statistics like maximum, minimum, mean*

4.2 GNI

The Global Financial Database groups countries into 4 categories based on their gross national income (GNI). Countries with less than $1,035 GNI per capita are classified as low income economies, those with GNI per capita between $1,036 and $4,085 as lower middle income economies, those with GNI per capita between $4,086 and $12,615 as upper middle income economies and finally those with GNI per capital of $12,616 or higher as high income countries. We have generated summary statistics for different indicators and grouped them on the basis of their GNI per capital. Figures 7 to 12 represent summary statistics of percentage of people: owning an account, owning an account at a financial institution, owning a card, using a card, having internet access, and owning a phone respectively.

% Account (age 15+)

Income <fctr>	Count <dbl>	Min <dbl>	Max <dbl>	Mean <dbl>	Median <dbl>
Low income	2	43.50099	49.49275	46.49687	46.49687
Lower middle income	7	26.03280	83.56481	51.28014	47.79103
Upper middle income	7	44.17417	88.70902	70.16256	71.62709
High income	10	85.74219	99.97659	97.36625	99.03776

Fig. 7: *Percentage of people owning an account grouped by GNI*

% Financial Institution Account (age 15+)

Income <fctr>	Count <dbl>	Min <dbl>	Max <dbl>	Mean <dbl>	Median <dbl>
Low income	2	28.44291	38.63492	33.53892	33.53892
Lower middle income	7	23.19142	83.56481	44.46602	35.71186
Upper middle income	7	44.17417	88.70902	68.44107	66.32544
High income	10	84.56073	99.97659	97.21207	99.03776

Fig. 8: *Percentage of people owning an account at a financial institution grouped by GNI*

From the statistics above, we found that in high income economies and upper middle economies, the percentage of people owning an account is very similar to the percentage of people owning an account at a financial institution. This implies that most people in these economies have and actively use a bank account. However, this is not the case in the other economies. On average, only 33% of people own a financial institution account compared to 46% of people owning an account in low income economies. This means that around 30% of the people that own an account have some sort of mobile money account or another form of digitalized banking.

% Owns Credit/Debit Card (age 15+)

Income <fctr>	Count <dbl>	Min <dbl>	Max <dbl>	Mean <dbl>	Median <dbl>
Low income	2	14.55933	21.87676	18.21804	18.21804
Lower middle income	7	11.99620	72.37830	28.35388	22.13079
Upper middle income	7	27.09842	77.54442	51.36669	56.47530
High income	10	72.36775	98.12654	93.17196	96.84606

Fig. 9: *Percentage of people owning a credit/debit card grouped by GNI*

% Used Credit/Debit Card (age 15+)

Income <fctr>	Count <dbl>	Min <dbl>	Max <dbl>	Mean <dbl>	Median <dbl>
Low income	2	6.912231	12.01185	9.46204	9.46204
Lower middle income	7	5.369735	56.52353	15.43303	8.26949
Upper middle income	7	8.430479	55.26465	35.94119	43.52207
High income	10	60.996938	96.71321	87.14491	92.01893

Fig. 10: *Percentage of people owning a credit/debit card grouped by GNI*

There is a significant disparity between credit/debit card ownership and usage in high income economies (93.17% and 87.14%) compared to low income economies (18.21% and 9.46%). This could be due to digitalized payment systems being more accessible and convenient in low income economies. However, there is still a long way to go for this disparity to be addressed.

% Internet access (age 15+)

Income <fctr>	Count <dbl>	Min <dbl>	Max <dbl>	Mean <dbl>	Median <dbl>
Low income	2	20.38413	25.74126	23.06270	23.06270
Lower middle income	7	27.57922	89.29688	53.05965	49.97822
Upper middle income	7	61.11693	85.16717	73.84284	78.13150
High income	10	75.61539	98.88784	90.71548	93.79902

Fig. 11: *Percentage of people having access to the internet grouped by GNI*

% Owns Phone (age 15+)

Income <fctr>	Count <dbl>	Min <dbl>	Max <dbl>	Mean <dbl>	Median <dbl>
Low income	2	54.59597	70.59063	62.59330	62.59330
Lower middle income	7	65.56514	98.89006	82.84456	80.97563
Upper middle income	7	84.66303	100.00000	90.44164	90.60962
High income	10	88.23817	100.00000	95.33132	96.13325

Fig. 12: *Percentage of people owning a phone grouped by GNI*

Unlike most other indicators analyzed here, there is not a huge amount of difference between phone ownership rates in low income economies compared to phone ownership rates in high income economies.

4.3 Continent

The next set of figures (13 to 15) are similar summary statistics but are grouped by continents instead. This was done using the new continent column that was added to the dataset as mentioned previously in the report.

% Account (age 15+)

Continent <chr>	Count <dbl>	Min <dbl>	Max <dbl>	Mean <dbl>	Median <dbl>
Africa	6	27.44401	85.37811	50.40250	46.49687
Americas	7	26.03280	99.63462	70.49961	71.62709
Asia	6	47.79103	98.48942	82.63511	87.22561
Europe	5	44.17417	99.97659	85.39133	99.47777
Oceania	2	98.75238	99.32315	99.03776	99.03776

% Financial Institution Account (age 15+)

Continent <chr>	Count <dbl>	Min <dbl>	Max <dbl>	Mean <dbl>	Median <dbl>
Africa	6	26.09891	84.11220	40.37508	32.48078
Americas	7	23.19142	99.63462	68.55305	66.32544
Asia	6	36.14336	98.48942	80.39901	86.63487
Europe	5	44.17417	99.97659	85.39133	99.47777
Oceania	2	98.75238	99.32315	99.03776	99.03776

Fig. 13: *Percentage of people owning an account, financial institution account grouped by Continent*

% Owns Credit/Debit Card (age 15+)

Continent <chr>	Count <dbl>	Min <dbl>	Max <dbl>	Mean <dbl>	Median <dbl>
Africa	6	14.55933	60.37756	26.20353	21.18531
Americas	7	11.99620	97.92484	56.54665	56.47530
Asia	6	25.86440	93.71634	64.58675	74.95609
Europe	5	27.09842	98.12654	78.29808	96.68455
Oceania	2	97.00758	97.13351	97.07054	97.07054

% Used Credit/Debit Card(age 15+)

Continent <chr>	Count <dbl>	Min <dbl>	Max <dbl>	Mean <dbl>	Median <dbl>
Africa	6	5.568744	49.06664	15.41270	9.458384
Americas	7	5.369735	96.71321	47.00079	43.522075
Asia	6	8.269490	83.93211	48.95362	58.111915
Europe	5	8.430479	96.30330	69.05903	90.697896
Oceania	2	94.388384	95.10569	94.74704	94.747037

Fig. 14: *Percentage of people owning/using a credit/debit card grouped by Continent*

% Internet Access (age 15+)

Continent <chr>	Count <dbl>	Min <dbl>	Max <dbl>	Mean <dbl>	Median <dbl>
Africa	6	20.38413	64.22569	40.19061	34.21654
Americas	7	49.97822	94.80496	75.82533	78.29549
Asia	6	27.57922	93.80952	72.35367	78.00077
Europe	5	78.13150	98.88784	89.45076	89.29688
Oceania	2	93.78852	94.51241	94.15047	94.15047

% Owns Phone (age 15+)

Continent <chr>	Count <dbl>	Min <dbl>	Max <dbl>	Mean <dbl>	Median <dbl>
Africa	6	54.59597	87.75913	77.06770	80.94082
Americas	7	69.82634	97.18563	87.13898	88.23817
Asia	6	65.56514	100.00000	92.22791	96.48890
Europe	5	90.60962	100.00000	94.66521	92.24911
Oceania	2	95.45861	96.97202	96.21532	96.21532

Fig. 15: *Percentage of people with internet access, and owning a phone grouped by Continent*

Africa lags behind other continents for every indicator except phone ownership (which is pretty much the same). For example, credit/debit card usage rates in Africa is only 15.41% whereas it is almost 3x as much in North/South America and Asia. This could be because the vast majority of countries in Africa are classified as low income or lower middle income which means there is some disparity between Africa and other continents in the dataset.

5 Data Analysis

5.1 Global

We performed further data analysis using data from all the countries (~140) across 4 years and plotted them using the “ggplot2” package in R. A lot of the plots are self-explanatory and also help with visualization of the information contained in the dataset.

5.1.1 Account Ownership

From Fig. 16 below, we can see that on average, the percentage of people having an account in low income countries increased from ~10% in 2011 to ~43% in 2021. Furthermore, there is a huge gap between account ownership rates in low income countries like Mali, Mozambique and Yemen compared to high income countries such as Canada, USA or Norway.

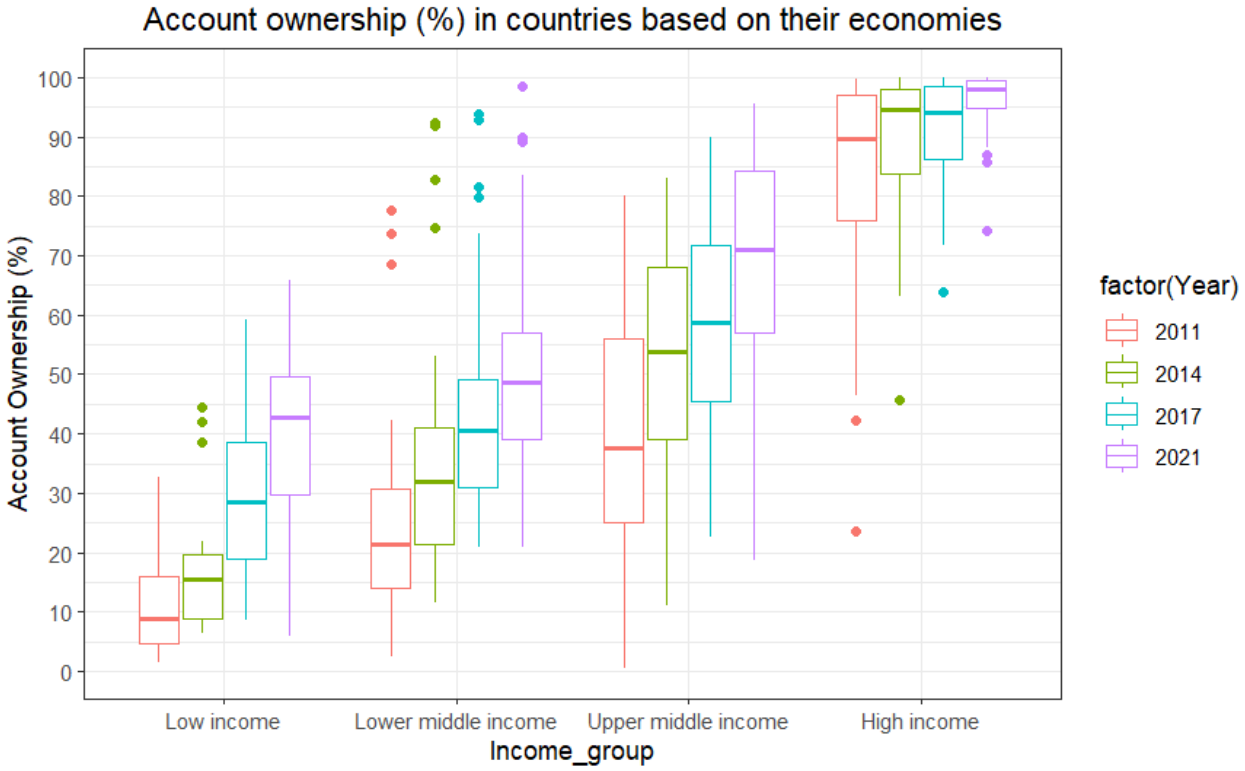

Fig. 16: *Account Ownership rates based on GNI*

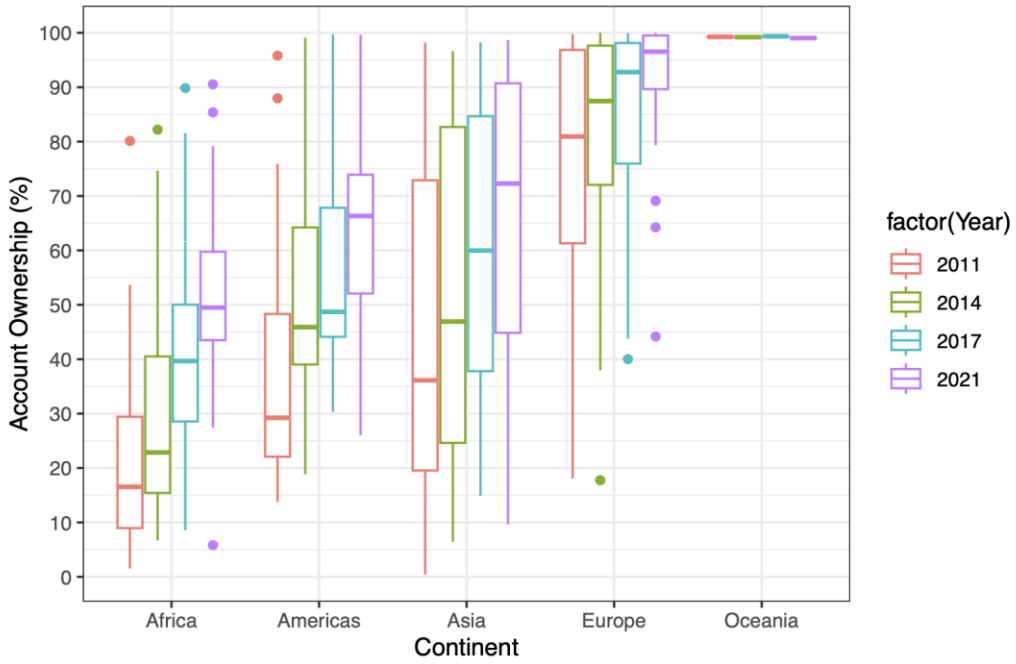

Fig. 17: *Account Ownership rates based on Continent*

5.2.2 Financial Institution Account Ownership

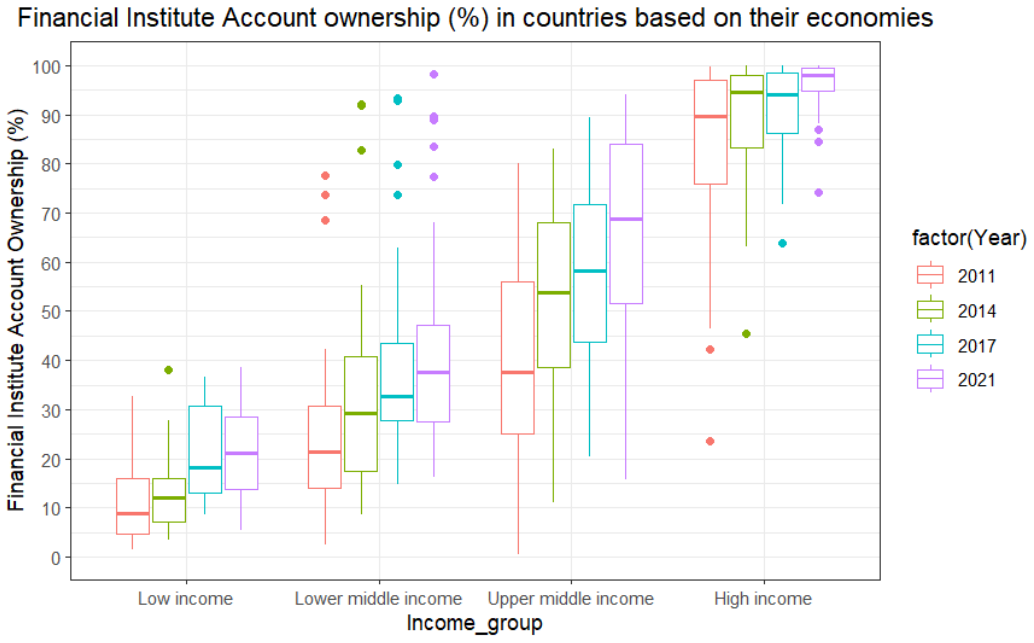

Fig. 18: *Financial Institution Account Ownership rates based on GNI*

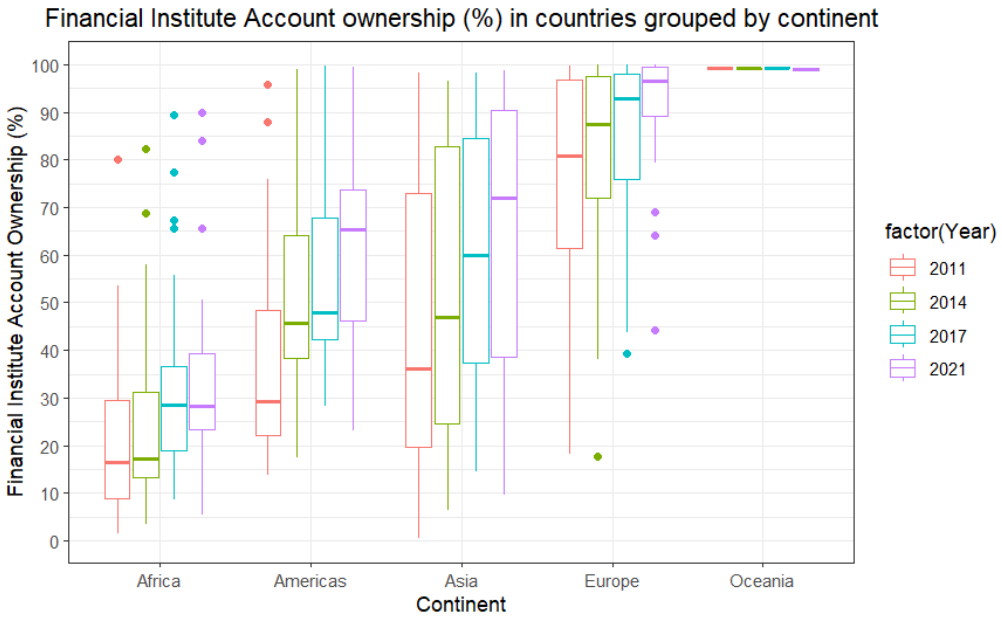

Fig. 19: *Financial Institution Account Ownership rates based on Continent*

Account ownership rates in Oceania (Fig. 19) are really high across all 4 years because it only Australia and New Zealand actually have data, and they are both high income economies. Countries in Africa primarily consist of low income and lower middle income economies and have far lesser people owning an account compared to Europe or Oceania.

5.2.3 Credit/Debit Card Ownership

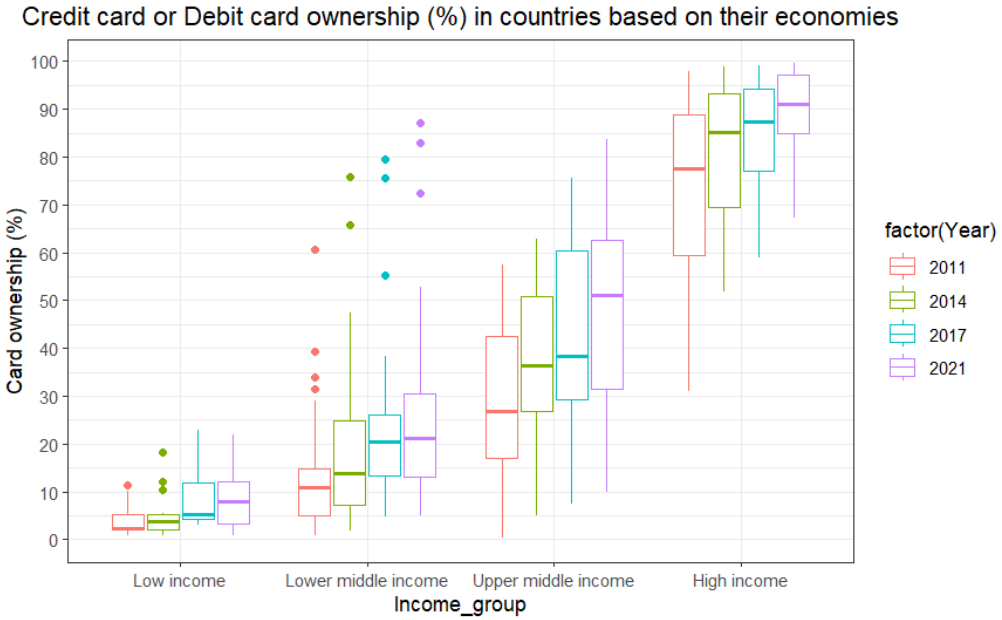

Fig. 20: *Credit/Debit Card Ownership rates based on GNI*

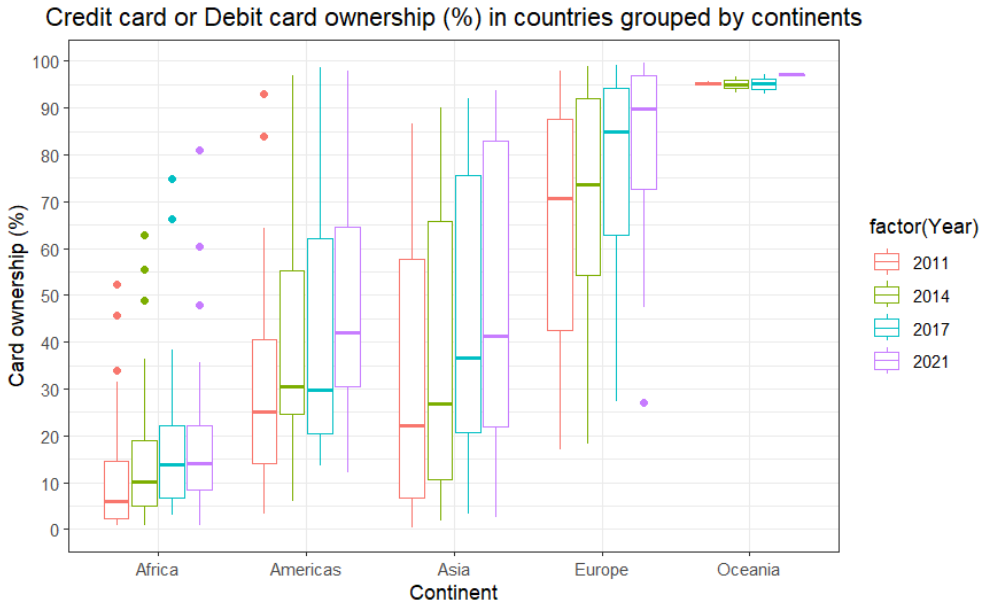

Fig. 21: *Credit/Debit Card Ownership rates based on Continent*

5.2 United States

We also decided to perform some analysis with just US data and compare it to the data from the rest of the world. The US is classified as a high income economy, so it is no surprise that it has really high percentages for all the indicators that we use in our analysis. Additionally, we also looked at US data over the last decade to try and identify any patterns or common trends.

Figure 21 shows the percentage of people owning an account in the United States across time. Over the last decade, this number has gone up from ~88% to ~95%, indicating that the US is on the right path in terms of financial inclusion. Figure 22 shows how the US compares with respect to the global averages of different income groups (GNI). We notice that the account ownership rates in the US are on the higher end across all 4 years.

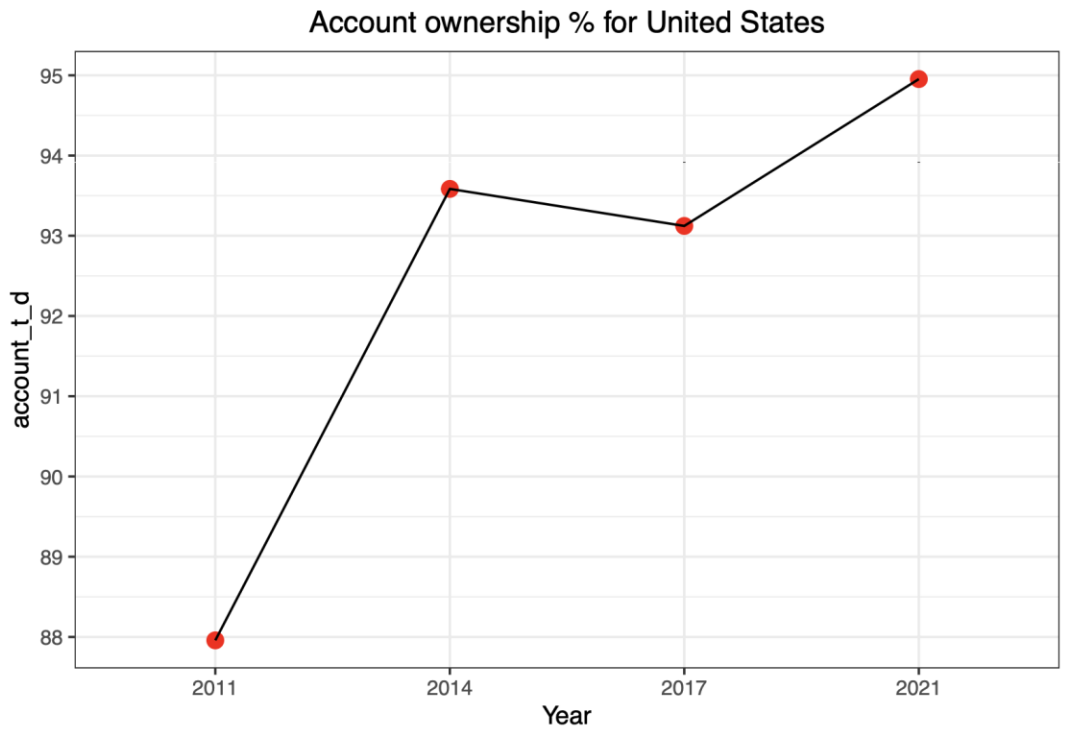

Fig. 21: *Account Ownership rates in US from 2011 to 2021*

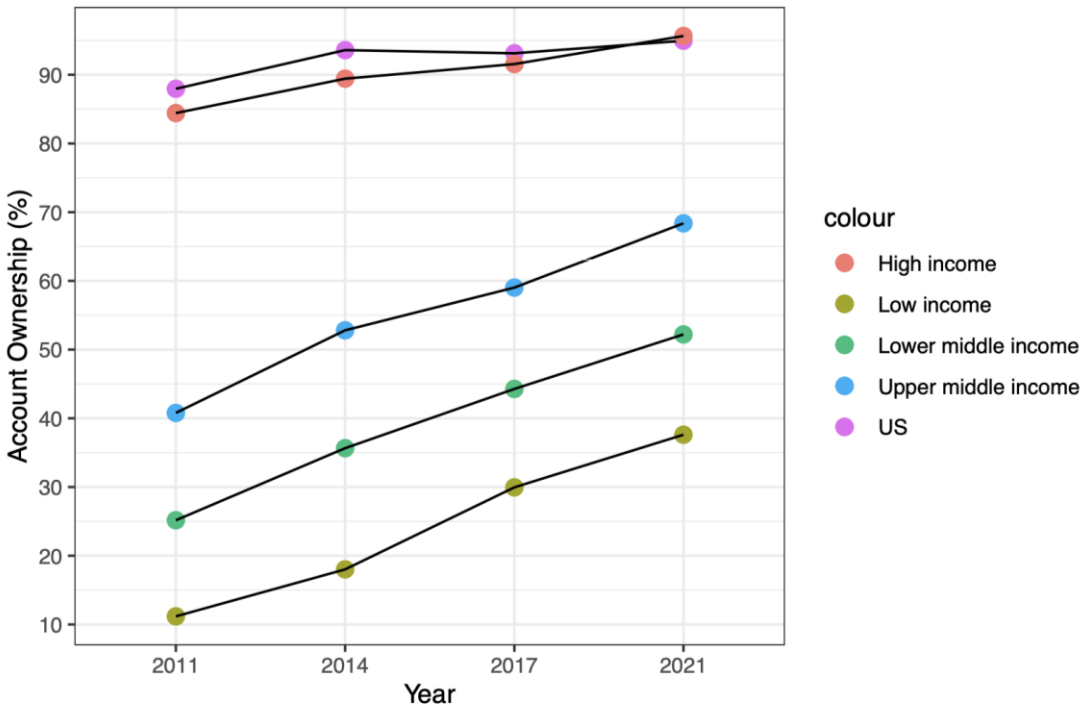

Fig. 22: *US vs Global Account Ownership rates from 2011 to 2021*

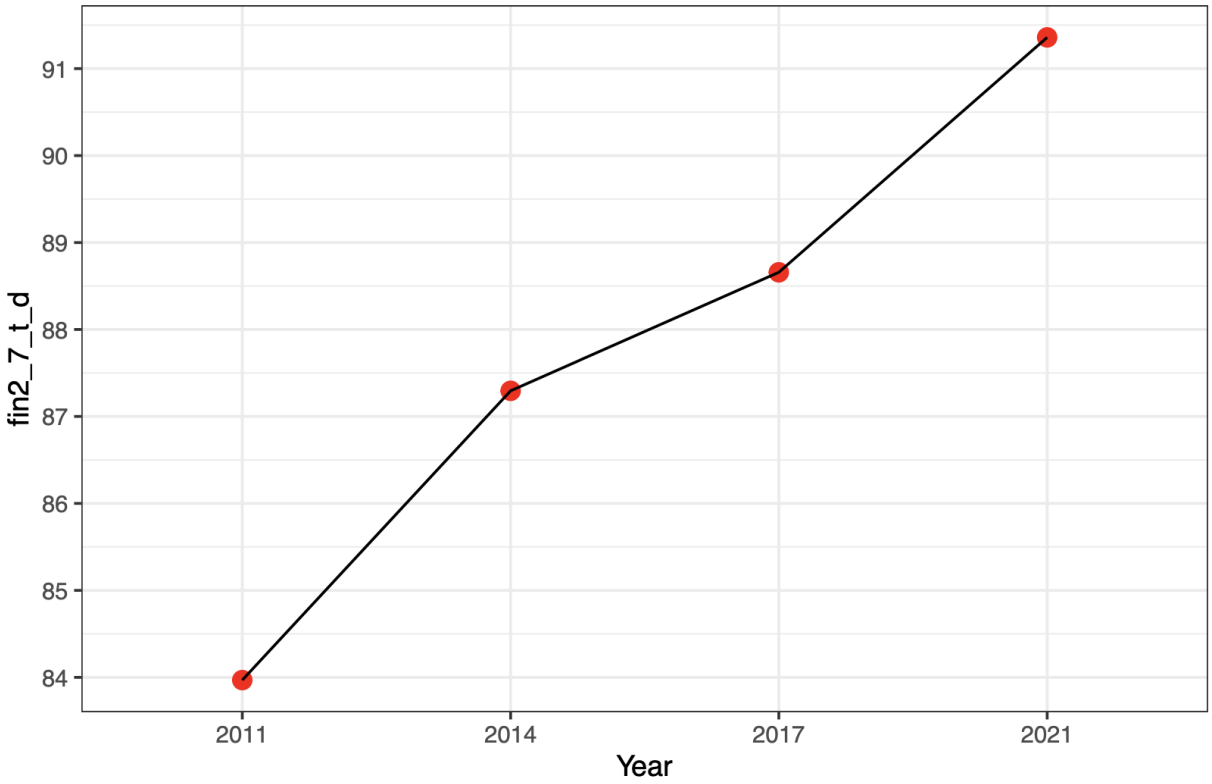

Fig. 23: *Credit/Debit card ownership rates in US from 2011 to 2021*

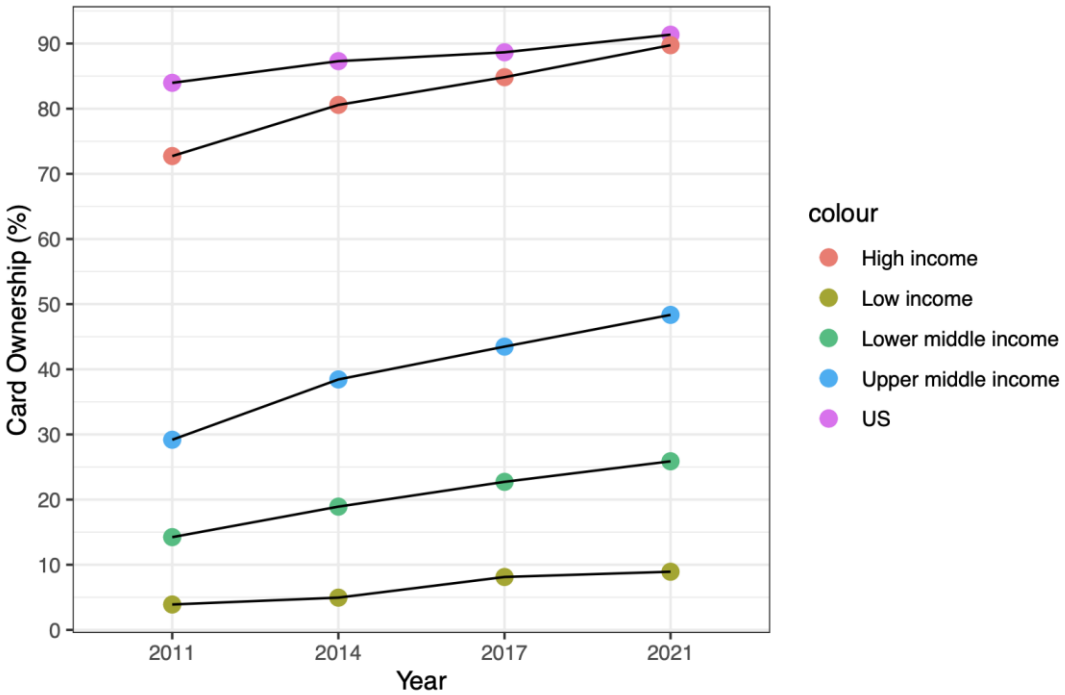

Fig. 24: *US vs Global Credit/Debit card ownership rates from 2011 to 2021*

6 Machine Learning Models

6.1 KNN

KNN stands for k-nearest neighbors, and it is a machine learning algorithm used for classification and regression tasks. In our case, we are majorly using it for classification of data. KNN is a non-parametric method, which means it does not make any assumptions about the underlying distribution of the data. We have chosen the K values after a lot optimizing and fine-tuned them based on our selected variables. We have narrowed down the KNN data set to six columns that we believe are the most relevant for predicting data: continent, year, income group having an account, having a financial institution account, owns a credit or debit card. By focusing on these specific features, we can reduce the complexity of the data set and potentially improve the accuracy of our KNN predictions. Figure 25 below represents the summary statistics for our KNN model.

```
  Continent           Year                 Income_group    Account         Fin_Inst_Account     Owns_card
 Length:546         2011:141   Low income         : 67   Min.   :  0.4049   Min.   :  0.4049   Min.   : 0.2695
 Class :character   2014:140   Lower middle income:155   1st Qu.: 31.9050   1st Qu.: 27.3048   1st Qu.:13.6644
 Mode  :character   2017:144   Upper middle income:147   Median : 55.9124   Median : 52.4660   Median :35.4781
                    2021:121   High income        :177   Mean   : 57.8583   Mean   : 55.4761   Mean   :43.7204
                                                         3rd Qu.: 88.2167   3rd Qu.: 88.1135   3rd Qu.:74.6390
                                                         Max.   :100.0000   Max.   :100.0000   Max.   :99.6180
```

Fig. 25: *Summary statistics for the KNN dataset*

6.1.1 GNI

Prediction	Reference			
	Low income	Lower middle income	Upper middle income	High income
Low income	12	10	2	0
Lower middle income	6	24	13	1
Upper middle income	0	8	24	7
High income	0	2	5	50

Fig. 26: *Confusion matrix of KNN analysis (GNI group)*

The confusion matrix has four categories or classes - low income, lower middle income, upper middle income and high income. The rows represent the predicted values, while the columns represent the actual values. The numbers in the cells show the number of instances that were predicted to belong to a particular class and the actual class that they belong to. For instance, in this case, 12 instances were predicted to be low income and actually belonged to Low income, while 6 instances were predicted to be lower middle income but actually belonged to low income.

The diagonal cells represent the correctly classified instances, while the off-diagonal cells represent the misclassified instances. Looking at the matrix, we can see that the model performs well at correctly classifying high income instances, with 50 out of 57 being correctly classified. However, the model struggles more with the other income groups, with varying levels of misclassification in each.

```
                  Accuracy : 0.6707
                    95% CI : (0.5932, 0.742)
       No Information Rate : 0.3537
       P-Value [Acc > NIR] : < 2.2e-16

                     Kappa : 0.546

    Mcnemar's Test P-Value : NA
```

Fig. 27 : *Overall statistics of KNN analysis (GNI group)*

The accuracy of the model is 0.6707, which means that 67.07% of the predictions made by the model are correct. The 95% confidence interval (CI) for the accuracy of the model is (0.5932, 0.742), which suggests that we can be 95% confident that the true accuracy of the model falls within this range. The "No Information Rate" is 0.3537, which means that if we simply guessed the most common income group for all observations, we would be correct 35.37% of the time.

The p-value for the accuracy of the model being greater than the "No Information Rate" is very close to zero, which suggests that the model is significantly better than simply guessing the most common income group. The Kappa value is 0.546, which measures the agreement between the model's predictions and the actual income group. A kappa value of 1 indicates perfect agreement, and a value of 0 indicates agreement by chance alone. Mcnemar's Test P-Value is NA, which suggests that there is no significant difference between the model's predictions for different income groups.

	Class: Low income	Class: Lower middle income	Class: Upper middle income	Class: High income
Sensitivity	0.66667	0.5455	0.5455	0.8621
Specificity	0.91781	0.8333	0.8750	0.9340
Pos Pred Value	0.50000	0.5455	0.6154	0.8772
Neg Pred Value	0.95714	0.8333	0.8400	0.9252
Prevalence	0.10976	0.2683	0.2683	0.3537
Detection Rate	0.07317	0.1463	0.1463	0.3049
Detection Prevalence	0.14634	0.2683	0.2378	0.3476
Balanced Accuracy	0.79224	0.6894	0.7102	0.8980

Fig. 28: *Statistics by class for KNN analysis (GNI group)*

Below, we provide a breakdown of what the different statistics mean for the reader's understanding.

Sensitivity: It measures the proportion of actual positive cases that are correctly identified by the model. For example, for the low-income class, the sensitivity is 0.66667, which means that 66.67% of the true low-income individuals are correctly classified as low-income.

Specificity: It measures the proportion of actual negative cases that are correctly identified by the model. For example, for the low-income class, the specificity is 0.91781, which means that 91.78% of the true non-low-income individuals are correctly classified as non-low-income.

Positive Predictive Value (PPV): It measures the proportion of positive predictions that are true positive cases. For example, for the low-income class, the PPV is 0.5, which means that only 50% of the predicted low-income cases are actually Low income.

Negative Predictive Value (NPV): It measures the proportion of negative predictions that are true negative cases. For example, for the Low-income class, the NPV is 0.95714, which means that 95.71% of the predicted non-low-income cases are actually non-Low income.

Prevalence: It measures the proportion of the population that belongs to a certain class. For example, the prevalence of the high-income class is 0.3537, which means that 35.37% of the population belongs to the high-income class.

Detection Rate: It measures the proportion of actual cases that are correctly identified by the model. For example, for the low-income class, the detection rate is 0.07317, which means that the model correctly identifies 7.32% of the true low-income cases.

Detection Prevalence: It measures the proportion of predicted cases that belong to a certain class. For example, the detection prevalence of the high-income class is 0.3476, which means that 34.76% of the predicted cases belong to the high-income class.

Balanced Accuracy: It is the average of sensitivity and specificity, and it is a measure of overall model performance. For example, the balanced accuracy for the low-income class is 0.79224, which means that the model correctly identifies 79.22% of both Low and non-low-income cases.

Overall, the statistics for the high-income class are better followed by the low-income class compared to other classes.

6.1.2 Low Income Economies

Prediction	Reference: Low income	Lower middle income	Upper middle income	High income
Low income	12	0	0	0
Lower middle income	6	0	0	0
Upper middle income	0	0	0	0
High income	0	0	0	0

Fig. 29: *Confusion matrix for KNN analysis of low-income group*

In this case, the model did not make any predictions for upper middle income or high-income categories, and only predicted Low income and lower middle-income categories. There were 12 instances where the model predicted Low income and the actual value was also Low income.

```
Overall Statistics

               Accuracy : 0.6667
                 95% CI : (0.4099, 0.8666)
    No Information Rate : 1
    P-Value [Acc > NIR] : 1

                  Kappa : 0

 Mcnemar's Test P-Value : NA
```

Fig. 30: *Overall statistics for KNN analysis of low-income group*

The overall accuracy of the model is 0.6667, which means it correctly classified 66.67% of the observations. The 95% confidence interval for accuracy ranges from 0.4099 to 0.8666. The no-information rate is 1, which means if the model always predicted the most frequent category (in this case, Low income), it would have an accuracy of 1.

Statistics by Class:

	Class: Low income	Class: Lower middle income	Class: Upper middle income	Class: High income
Sensitivity	0.6667	NA	NA	NA
Specificity	NA	0.6667	1	1
Pos Pred Value	NA	NA	NA	NA
Neg Pred Value	NA	NA	NA	NA
Prevalence	1.0000	0.0000	0	0
Detection Rate	0.6667	0.0000	0	0
Detection Prevalence	0.6667	0.3333	0	0
Balanced Accuracy	NA	NA	NA	NA

Fig. 31: *Statistics by class for KNN analysis of low-income group*

The model's sensitivity is 0.6667 for low-income, and NA (not applicable) for the other categories since the model did not make any predictions for them. The specificity is NA for low-income since it did not predict any other category as low-income, and it is 0.6667 for lower middle income and 1 for upper middle income and high income. Prevalence is the proportion of the actual category in the dataset. In this case, the prevalence is 1 for Low income and 0 for the other categories. The detection rate is 0.6667 for low income and 0 for the other categories. The detection prevalence is 0.6667 for low-income and 0 for the other categories and the balanced accuracy is NA since the model only predicted 2 out of 4 categories.

6.1.3 High Income Economies

	Reference			
Prediction	Low income	Lower middle income	Upper middle income	High income
Low income	0	0	0	0
Lower middle income	0	0	0	1
Upper middle income	0	0	0	7
High income	0	0	0	50

Fig. 32: *Confusion matrix for KNN analysis of high-income group*

In this case, the model did not predict any labels in the low income and lower middle-income classes, and therefore the sensitivity, positive predictive value, and negative predictive value could not be calculated for these classes. The model predicted 7 instances in the upper middle-income class to be high income, resulting in a high number of false positives in that class. The model performed well in predicting the high-income class, with 50 true positives and no false negatives.

```
                Accuracy : 0.8621
                  95% CI : (0.7462, 0.9385)
    No Information Rate : 1
    P-Value [Acc > NIR] : 1

                   Kappa : 0

Mcnemar's Test P-Value : NA
```

Fig. 33: *Overall statistics for KNN analysis of high-income group*

The overall accuracy of the model is 0.8621, which means that the model correctly predicted the class label for 86.21% of the test instances. The no-information rate, which is the accuracy of predicting the most frequent class label, is 1, indicating that the model performed better than predicting the most frequent class label.

	Class: Low income	Class: Lower middle income	Class: Upper middle income	Class: High income
Sensitivity	NA	NA	NA	0.8621
Specificity	1	0.98276	0.8793	NA
Pos Pred Value	NA	NA	NA	NA
Neg Pred Value	NA	NA	NA	NA
Prevalence	0	0.00000	0.0000	1.0000
Detection Rate	0	0.00000	0.0000	0.8621
Detection Prevalence	0	0.01724	0.1207	0.8621
Balanced Accuracy	NA	NA	NA	NA

Fig. 34: *Statistics by class for KNN analysis of high-income group*

In this case, because the model did not predict any instances in the low-income and lower middle-income classes, the sensitivity, positive predictive value, and negative predictive value could not be calculated for these classes. The model had a high specificity for the low-income class and a high specificity for the lower middle-income class, indicating that it correctly identified all instances that were not in these classes. The model had a low specificity for the upper middle-income class, indicating that it incorrectly identified many instances in that class as high-income. Because the model did not make any errors, the positive predictive value and negative predictive value could not be calculated for any class, and the prevalence and detection prevalence are equal to the sensitivity and detection rate, respectively, for the high-income class.

6.2 Decision Tree

A Decision Tree model is a supervised machine learning algorithm that is used for both regression and classification tasks. The Decision Tree algorithm builds a tree-like model of decisions and their possible consequences. It splits the data into subsets based on the values of input features and recursively partitions the data into smaller and more homogeneous groups. The splits are determined by selecting the features that provide the most information gain or the best split based on a specific criterion like minimizing the mean squared error. This type of model can be useful in predicting outcomes based on multiple variables.

6.2.2 GNI

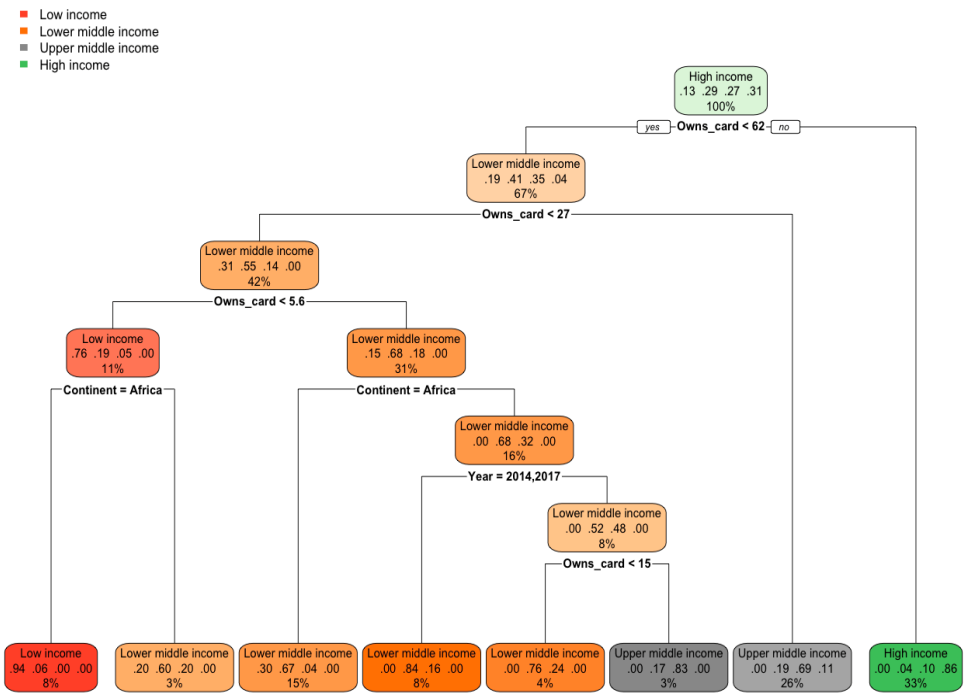

Fig 35: *Decision tree for income groups*

```
       17) Continent=Americas,Asia 10   4 Lower middle income (0.20000000 0.60000000 0.20000000
0.00000000) *
      9) Owns_card>=5.642907 117  38 Lower middle income (0.14529915 0.67521368 0.17948718
0.00000000)
       18) Continent=Africa 57  19 Lower middle income (0.29824561 0.66666667 0.03508772
0.00000000) *
       19) Continent=Americas,Asia,Europe 60  19 Lower middle income (0.00000000 0.68333333
0.31666667 0.00000000)
         38) Year=2014,2017 31   5 Lower middle income (0.00000000 0.83870968 0.16129032
0.00000000) *
         39) Year=2011,2021 29  14 Lower middle income (0.00000000 0.51724138 0.48275862
0.00000000)
           78) Owns_card< 15.3741 17   4 Lower middle income (0.00000000 0.76470588 0.23529412
0.00000000) *
           79) Owns_card>=15.3741 12   2 Upper middle income (0.00000000 0.16666667 0.83333333
0.00000000) *
     5) Owns_card>=26.76601 98  30 Upper middle income (0.00000000 0.19387755 0.69387755
0.11224490) *
   3) Owns_card>=61.97015 125  17 High income (0.00000000 0.04000000 0.09600000 0.86400000) *
```

Fig 36: *Decision tree for income groups (summary of the nodes)*

predictions	Low income	Lower middle income	Upper middle income	High income
Low income	8	1	0	0
Lower middle income	10	30	10	0
Upper middle income	0	11	27	9
High income	0	2	7	49

Fig 37: *Confusion matrix for the DTM based on income groups*

From the confusion matrix for income groups based on the decision tree model, we can see that the model is classifying all four income groups quite well. There are deviations but most of the observations are assigned to the nearest available category, which is a positive indicator.

6.2.3 Continents

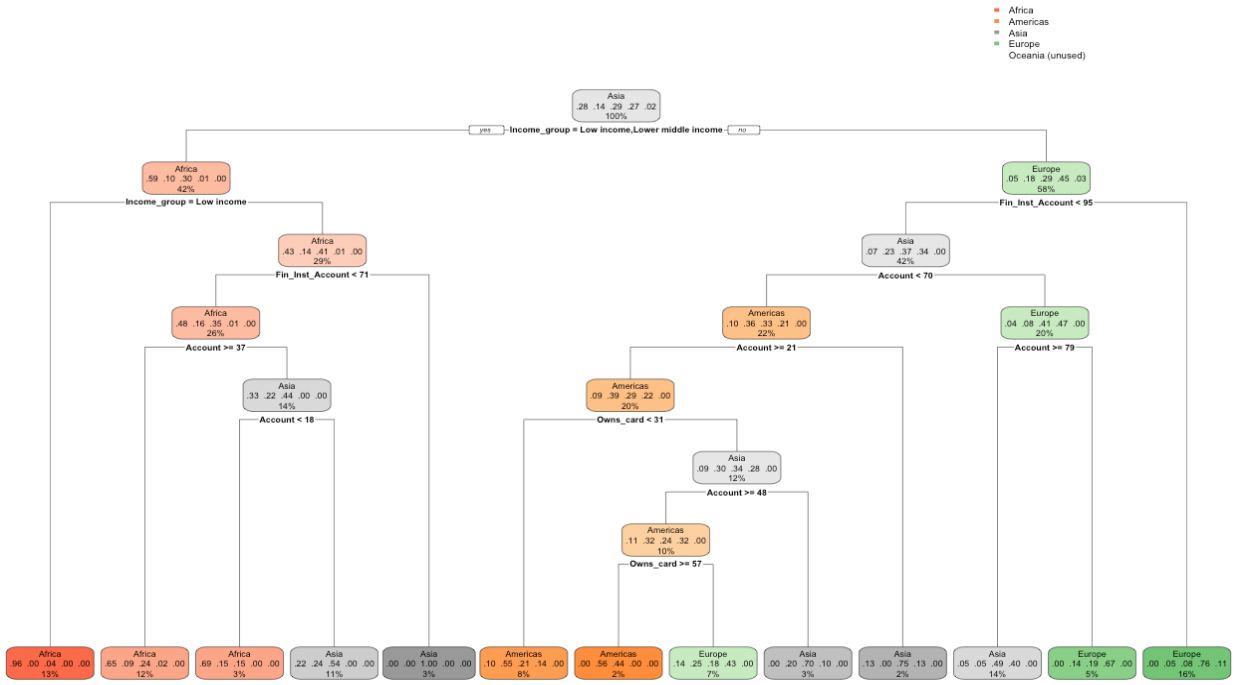

Fig 38: *Decision tree for continents*

Clearly this decision tree Is overfitted which means that it doesn't describe the data accurately. The topmost split is based on the variable income-group, which has two possible values: lower middle income and low Income based on Asia. If income group is lower middle income, the model looks at two other variables, continent and owns card, to make further splits. If Continent is either Americas or Asia, and owns card is less than 26.76601, the data point is classified as lower middle income and belongs to the

respective continent. If Continent is Africa and owns card is less than 5.642907, the data point is classified as lower middle income and belongs to Africa.If Continent is Americas, Asia, and Europe and owns card is either greater than or equal to 15.3741 or year is either 2014 or 2017, the data point is classified as lower middle income and belongs to the respective continent. If owns card is greater than or equal to 61.97015, the data point is classified as high income and belongs to Europe and so on until the tree reaches an end.

predictions	Africa	Americas	Asia	Europe	Oceania
Africa	21	2	16	2	0
Americas	3	6	5	1	0
Asia	12	11	17	18	0
Europe	2	9	13	25	1
Oceania	0	0	0	0	0

Fig 39: *Confusion matrix for the DTM based on continents*

The confusion matrix for continents is not that great as we can clearly observe a lot of deviations from the true category. This can be observed especially in Asia and Americas. Oceania has only 1 value, which is classified under Europe.

7 Conclusion

In summary, the analysis of the dataset with financial inclusion data shed light on various aspects of financial inclusion rates and socioeconomic factors across countries. The correlation matrix and summary statistics revealed important relationships and disparities between different indicators, such as account ownership, card usage, internet access, and phone ownership. The findings highlighted the need for further efforts to address disparities and promote financial inclusion, particularly in lower-income economies.

In our analysis, we used both the KNN and Decision Tree models to classify data based on income groups and continents. The accuracy of the KNN model ranged from 66.67% to 86.21%, depending on the specific income group or continent being considered. We also attempted KNN analysis using the variables of continent and year, but the accuracy rates were consistently lower than 50%, even after adjusting K-values.

In addition to KNN, we applied a Decision Tree model to the same dataset, which provided alternative insights into the classification task. The confusion matrices generated by both models showed how well they performed for each category, highlighting areas of accuracy as well as deviations from the true classifications.

The KNN model and Decision Tree model offered different perspectives and approaches to the classification task. While KNN relied on finding similar neighbors, the Decision Tree model utilized a hierarchical structure of decisions to make predictions. Among the income groups, the KNN analysis based on high-income groups achieved the highest overall accuracy of 86.21%. However, the Decision Tree model seemed to classify the data better overall compared to the KNN model.

It is important to note that the complexity of the dataset and the presence of missing data make it challenging to achieve high accuracy scores with any machine learning model. At present, we believe that data analysis provides a better method for identifying trends and patterns in financial inclusion rates. However, there is still much progress to be made before machine learning models can accurately predict financial inclusion rates in countries.

8 References

https://databank.worldbank.org/source/global-financial-inclusion
https://microdata.worldbank.org/index.php/catalog/4607/study-description
https://www.worldbank.org/en/publication/globalfindex/Report

9 Team Member Contributions

All of our team members made significant contributions to different aspects of the project. Rishabh Kansal and Yeshwanth Pendyala played a crucial role in data cleaning, ensuring the accuracy and reliability of the dataset. Sri Krishna Yerramilli conducted the exploratory data analysis, delving into the dataset to uncover patterns and relationships. Jahnavi Ravi focused on developing machine learning models to enhance the analysis and generate predictive insights. Additionally, Sri Krishna and Jahnavi assisted in creating visually appealing slides for the presentation and compiled the report. Each team member's expertise and effort were instrumental in the successful execution of the project, resulting in comprehensive analysis and valuable insights.