Master Thesis U.S.E.

Various types of Microcredits lead to different success rates: a study in sub-Saharan Africa

Author: Viënna van Holsteijn
Email: viennavanholsteijn@gmail.com
Date: 28th of June 2024
Student number: 4666240
Master: Sustainable Finance & Investments
Supervisor: Dr. Prof. Ronald Huisman
Second supervisor: Dr. Prof. Marie Dutordoir

ABSTRACT

This thesis investigates microcredit programs in Sub-Saharan Africa to discover their determinants of success. Investigating the historical trajectory of microfinance and existing microcredit models, this study identifies key success factors through multinomial logistic regression analysis and cluster analysis. Significant results have been found for differences in success rates between different microloans. The results of this research show that the type of microcredit with a higher success rate has a focus on individual loans, has a higher profit margin, and lower interest rates, which is in line with literature (Banerjee, 2013). In contrast, microcredits with lower success rates have a significantly lower focus on SME loans and a higher focus on individual loans, characterized by higher interest rates, which aligns with Banerjee (2013) that less successful microloans have higher interest rates. Interpreting these findings using existing literature, this research proposes implications for microfinance practice and future research. By gaining insights into possible factors determining the success rates of microcredits, this study contributes to targeted poverty alleviation and sustainable development strategies in Sub-Saharan Africa.

Keywords: Microcredit, Microfinance, Sub-Saharan Africa, Poverty Alleviation

JEL codes: G21, I32, O55

PREFACE

This paper contains the work of six (parttime) months of research, analysis and writing on microcredits in Sub-Saharan Africa. There has been much controversion about microcredits and their actual impact in alleviating poverty. Not much is known about the (current) success factors of successful microcredits, whereas this is an important question for both borrowers and lenders of microcredit.

The motivation behind this study stems from the need to provide a data-driven perspective on the impact of certain microcredit characteristics, offering insights into its effects on the success rate. By using a cluster analysis and a multinomial logistic regression analysis, this thesis aims to answer the central research question: *"Which types of microcredits are implemented in Sub-Saharan Africa, and how do these types of microcredits differ in terms of success?"*

Ultimately, this research aims to inform microcredit institutions and policy decisions regarding microloans and their implications for societal welfare. By shedding light on possible relations between microcredit factors and their success rates, this thesis intends to contribute to a more comprehensive understanding of the microcredit market, bridging the gap between academic research and real-world microcredit implications.

I would like to express my gratitude to my supervisor Dr. Prof. Ronald Huisman who has provided guidance, support, and insights throughout the process of completing this thesis. On top of that, Ronald has inspired me in his professional life through his own entrepreneural and academic activities. Besides, my family and in special Nils de Koning, has supported me through sharing tips in academic writing.

I hope that this thesis stimulates further discussion, encourages future research, and contributes to the ongoing efforts to create a fairer and more inclusive world for all members of society.

Viënna van Holsteijn
Utrecht University, Faculty of Law, Economics and Governance
28th of June 2024

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Motivation

This paper investigates whether different types of microcredits have a different success rate (equivalent to a low default rate) in microlending within sub-Saharan Africa. The exploration of microcredit programs in Sub-Saharan Africa is driven by a number of motivations deeply rooted in the region's socio-economic landscape and the broader global development agenda. Sub-Saharan Africa remains home to a significant proportion of the world's population living below the poverty line, with limited access to formal financial services worsening economic inequalities and hindering inclusive growth (World Bank, 2020). In this context, microcredit initiatives have emerged as an important tool for fostering financial inclusion and empowering communities within Africa and other poorer areas of the world.

The societal relevance of investigating microcredit programs in Sub-Saharan Africa extends beyond economic considerations, including social and environmental dimensions of well-being. By providing vulnerable populations, in particular women and rural dwellers, with access to microfinance services, these programs have the potential to unlock human capital, promote gender equality, and enhance resilience to shocks and vulnerabilities (Daley-Harris, 2009; Duflo, 2012). Moreover, microcredit interventions have been important in their role in promoting entrepreneurship, stimulating local economies, and contributing to the Sustainable Development Goals (SDGs) introduced by the United Nations (UN) (United Nations, 2015).
However, the realization of the positive potential of microcredit in Sub-Saharan Africa is contingent upon addressing many challenges and complexities inherent in the region's socio-economic context. Persistent poverty, political instability, and weak institutional capacity pose obstacles to the effective implementation and sustainability of microfinance initiatives (Armendáriz de Aghion & Morduch, 2006; Mersland & Strøm, 2009). Moreover, critiques regarding the commercialization of microfinance, high interest rates, over-indebtedness, and inadequate regulation underscore the great opportunities microcredit could have had (Bateman, 2010; Morduch, 1999).

In light of these considerations, the investigation of microloans in Sub-Saharan Africa is important to be able to make a significant positive impact on the people in Sub-Saharan Africa. The lower the success rates of microfinance, and thus the higher the default rate, the less people are helped where these microloans are meant for in the first place. This research can inform microfinance institutions as well as policymakers what different characteristics make certain microfinance more successful than others which creates both a positive impact for the lenders as well as the borrowers.

1.2 Academic relevance

The academic supply surrounding microcredit has undergone significant development in the previous years, reflecting both theoretical and empirical insights as a development tool. Early studies on microfinance showed its potential to help reducing poverty, empower marginalized populations, and increase economic growth (Hulme & Mosley, 1998; Johnson & Rogaly, 1997; Matin & Rutherford, 2002; Morduch, 1999). However, as microfinance gained importance as a mainstream development strategy, academics began to critically examine its underlying assumptions, implementation techniques, and impact (Armendáriz de Aghion & Morduch, 2006; Bateman, 2010).

Research within microcredit shows a diverse array of perspectives, methodologies, and thematic focuses, reflecting the multidimensional nature of microfinance. Academics have interrogated the complex interactions between microfinance and broader socio-economic processes, including gender dynamics, institutional arrangements, regulatory frameworks, and technological innovations (Daley-Harris, 2009; Duflo, 2012; Mersland & Strøm, 2009). Moreover, the emergence of randomized controlled trials (RCTs) and quasi-experimental methods has enabled researchers to evaluate the impact of microfinance interventions on various outcome variables, ranging from household income and expenditure patterns to women's empowerment and social capital (Banerjee et al., 2015; Duflo, 2012).

Despite the increase of empirical studies and theoretical frameworks, the academic discourse on microcredit remains characterized by ongoing debates and unresolved questions. Academics continue to have different views on issues such as over-indebtedness, client protection, sustainability, and the role of microfinance within broader development agendas (Bateman, 2010; Mersland & Strøm, 2009). Moreover, the contextual heterogeneity of microfinance operations across different regions and socio-economic contexts necessitates context-specific analyses and comparative assessments to discern patterns, trends, and best practices (Armendáriz de Aghion & Morduch, 2006; Mersland & Strøm, 2009).

The success rate of microcredits is a critical aspect that needs to be looked at in more detail. Microcredit programs, spearheaded by Microfinance Institutions (MFIs), have emerged as crucial tools for poverty alleviation and sustainable development. Success factors contributing to the effectiveness of microcredit models include innovative repayment structures, competitive interest rates, and diverse borrower groups (Mosley & Hulme, 1998; Banerjee, 2013; Morduch, 1999). Innovative repayment structures, such as those pioneered by the Grameen Bank, enable borrowers to repay loans through small, frequent installments, facilitating access to larger repeat loans. Competitive interest rates, particularly below 30% per year (Banerjee, 2013), are associated with successful microcredit programs. Diverse borrower groups and effective default control mechanisms also contribute to program success. However, much of the existing research on microcredits predates the 21st century (Hulme & Mosley, 1996; Hulme & Mosley, 1998; Johnson & Rogaly, 1997; Matin & Rutherford, 2002; Morduch, 1999), underscoring the need for updated insights and a reevaluation of success factors.

By critically evaluating data regarding microfinance instutions and their microfinance loans, this research aims to contribute to the ongoing academic discourse on microcredit, particularly within the context of Sub-Saharan Africa. By testing the determinants of success, identifying implementation challenges, and examining the differential impacts of microfinance interventions, this study seeks to enrich theoretical debates, inform policy decisions, and guide future research directions in the field of microfinance and development studies.

1.3 Research problem statement

The central focus of this research is to investigate the implementation of microcredits in Sub-Saharan Africa and discern how these various types of microcredits differ in terms of success. The primary research question guiding this study is:

"Which types of microcredits are implemented in Sub-Saharan Africa, and how do these types of microcredits differ in terms of success?"

In this research, the primary inquiry revolves around understanding the landscape of microcredit implementation in Sub-Saharan Africa and investigating the factors influencing their success. To comprehensively address this overarching research question, a series of sub questions will guide the investigation. Firstly, an examination of the current perspective on microcredits will be investigated with the first sub question: *What is the current perspective on microcredits? Is it indeed accurate to assert that it is negative regarding its outcomes?*

Following this, an exploration into the types of microcredits mentioned in academic literature, practiced in the field, and discussed online will be undertaken to grasp the diversity of microcredit models prevalent in the region. This will then be the second sub question: *Which types of microcredits are mentioned in the literature, practiced in the field, and discussed on the internet?*

Subsequently, a review of literature concerning the success factors of microcredits will shed light on the determinants of effective microcredit programs. This relates to the third sub question: *Is there literature concerning the success factors of microcredits, and if so, what are these factors?*

Lastly, a cluster analysis and multinomial logistic regression analysis will be performed to investigate whether specific types of microcredits exhibit higher success rates than others, thereby uncovering the distinguishing characteristics contributing to their success. This will be done with the fourth sub question: *Are specific types of microcredits more successful than others, and if yes, what are the distinguishing characteristics contributing to their success?*

Through this process, the study aims to offer a nuanced understanding of microcredit dynamics in Sub-Saharan Africa and provide insights to inform policy, practice, and future research.

This research seeks to contribute to the existing body of knowledge on microfinance by providing empirical insights into the implementation and effectiveness of microcredits in Sub-Saharan Africa. By addressing the research question and subquestions outlined above, this study aims to:

- Enhance understanding of the diverse landscape of microcredit models in Sub-Saharan Africa.
- Identify key success factors associated with microcredit loans.
- Provide practical recommendations for policymakers and practitioners, to optimize the impact of microfinance initiatives in the region.

The results of this research show that the type of microcredit with the higher success rate, has a focus on individual loans, has a higher profit margin and lower interest rates, which is in line with literature (Banerjee, 2013). The microcredit with a lower success rate, has a significant lower focus on SME loans, and a higher focus on individual loans and is characterized by higher interest rates. Which is also in line with Banerjee (2013), that less successful microloans have higher interest rates. Significant differences have been found between the different clusters and provide valuable information for microfinance institutions and policymakers.

The remainder of this research thesis is organized as follows: Section 2 describes the conceptual model and hypotheses to be tested, Section 3 outlines the data sources and methodology, and the variables descriptives, and Section 4 presents the results and interpretation of the cluster analysis and the multinomial logistic regression analysis. Section 5 concludes.

2. THEORY, LITERATURE REVIEW & HYPOTHESES

2.1 The emergence of microfinance

Microfinance has been a topic of interest for researchers, policy makers, corporations, and financial institutions worldwide. The evolution of microfinance has been marked by three distinct waves, each characterized by different approaches to poverty reduction (Matin et al., 2002).

The first wave was characterized by state-mediated and subsidized credit aimed at small farmers, believing this would increase productivity and incomes. The second wave emerged in the 1980s, where the focus shifted towards women running microenterprises. Small business loans were seen as tools for social empowerment and income generation. The third wave in the 2000s recognized the poor as a heterogeneous group with complex livelihoods and is also called the 'microfinance movement' (Hulme and Mosley, 1996; Johnson and Rogaly, 1997). Microfinance became a means to achieve household priorities, reduce vulnerability, and increase income (Matin et al., 2002).

Debates about finance and poverty-reduction have been shaped by changing conceptualizations of who the poor are and the nature of poverty. During the early development decades (1950s, 1960s and 1970s), the bulk of the poor were seen as the members of families headed by (male) small farmers. Their poverty could be overcome by subsidized agricultural credit that would raise productivity and incomes. From the early 1980s, a new image began to dominate thinking and action: the poor were mainly women (and their dependants) who coped with their situation by running microenterprises. Small business loans would permit them to expand (or establish) income-generating activities, raise their income, and socially empower them. Most recently, the poor have been conceptualized as a heterogeneous group of vulnerable households with complex livelihoods and varied needs. From such a perspective, microfinance is seen as a means for achieving household priorities (e.g. paying school fees, meeting funeral expenses), reducing vulnerability (e.g. a sudden drop in consumption, income, or assets), and/or increasing income (Matin et al., 2002).

Despite these evolutions, misconceptions about the financial markets of the poor persist. One of the most prevalent is that the poor do not and cannot save. However, research has shown that the poor do save, and that they do so in ways that are often more sophisticated than those of the non-poor (Matin et al., 2002).

2.2 Different types of microcredits

Different types of microcredits can be divided into three types: informal, semiformal, and formal (Matin et al., 2002). The traditional categories of ‘informal’ and ‘formal’ financial service providers have become more complex with the emergence of microfinance institutions (MFIs), which can be categorized as ‘semi-formal’. Informally, credit sources like money lenders, pawnbrokers, and traders were predominant, but this category also encompasses services from rotating and accumulating savings and credit associations, as well as deposit takers. Formal entities are recognized under the country’s banking regulations and offer standard retail services while participating in financial intermediation. MFIs, falling into the semi-formal category, often operate as registered NGOs or cooperatives and sometimes as specially chartered banks. Credit often serves as an insurance substitute in the realm of informal finance.

2.3 Success rate of microcredits

Microcredit programs, pioneered by Microfinance Institutions (MFIs), have become instrumental in addressing poverty and fostering sustainable development. This section comprehensively examines various success factors that contribute to the effectiveness of different microcredit models.

Microfinance has introduced innovative repayment structures, exemplified by the Grameen Bank's model, widely adopted by MFIs. This model allows borrowers to repay loans through small, frequent, and manageable installments, coupled with quick access to larger repeat loans. Such innovative structures have played a pivotal role in establishing the success of microcredit programs (Mosley & Hulme, 1998). Examining the correlation between repayment features and program success, studies by Mosley and Hulme (1998) highlight crucial factors such as the frequency of installments and repayment incentives. These incentives, including access to larger repeat loans and cash-back for timely repayments, contribute significantly to the overall success of microcredit initiatives.

Interest rates in microcredit programs exhibit considerable disparities globally. While some regions, such as Mexico and South Africa, face rates exceeding 100% per year, successful models in countries like Bangladesh, Bolivia, India, and Indonesia maintain significantly lower interest rates, often below 30% per year (Banerjee, 2013). Comparative studies with moneylender interest rates underscore the importance of competitive rates for program success.

A key aspect of microcredit success lies in the ability to serve diverse and unfamiliar borrower groups. In contrast to traditional moneylenders, who often have a limited clientele, MFIs extend their services to a broader demographic. Despite this diversity, successful MFIs maintain impressively low loan default rates, often below 10% and, in some cases, below 2% (Morduch, 1999).

Factors contributing to the lower interest rates and effective default control mechanisms in successful microcredit programs are multifaceted. These include innovative repayment structures, a diverse borrower demographic, and a correlation between repayment features and program success.

Building on these success factors, microcredit programs demonstrate broader impacts on poverty reduction, economic empowerment, and overall improvements in the well-being of beneficiaries. As a tool for social and economic development, microcredit's role extends beyond individual financial transactions to shaping broader social policies. Understanding the intricate linkages between microcredit and social policy is crucial for maximizing its positive impact on society.

In summary, the success of microcredit programs is rooted in innovative repayment structures, competitive interest rates, diverse borrower groups, and effective default control mechanisms. These success factors contribute not only to the financial sustainability of microcredit but also to its broader impact on poverty alleviation and sustainable development. However, most research has been done around or before the 21st century (Hulme & Mosley, 1996; Hulme & Mosley, 1998; Johnson & Rogaly, 1997; Matin & Rutherford, 2002; Morduch, 1999). Only one fairly recent research has been done in 2013, but this is still more than a decennium ago (Banerjee, 2013). Because of this time gap, it is important to reevaluate what defines successful microcredits and the differences in success rates between different microcredits. Besides, no current research exists around the focus on these microcredits, for example: microenterprises, SME’s, urban, rural or individual loans. This research will add knowledge about the correlation between success rates and the focus area of these loans.

2.4 Hypotheses development

Drawing from the literature review on microfinance and microcredits, the following hypotheses are developed to explore the relationship between different characteristics of microcredits and their success rates.

Our first hypothesis aims to assess whether different types of microcredits demonstrate varying success rates, particularly in terms of repayment performance. We want to test whether certain microcredit categories may have higher success rates than others.

Previous research suggests that microfinance has evolved through distinct waves, each targeting different segments of the population (Matin et al., 2002). The emergence of various microcredit models, including informal, semiformal, and formal institutions, shows the diversity within the sector (Matin et al., 2002). Additionally, studies have highlighted disparities in interest rates and default rates across regions and microfinance institutions (Banerjee, 2013; Morduch, 1999). By examining these variations, we seek to measure whether certain microcredit types exhibit superior performance in Sub-Saharan Africa. This reasoning leads us to the first hypothesis:

Hypothesis 1: *types of microcredits do not differ in success rates.*
Hypothesis alternative: *some types of microcredits have higher success rates than others.*

We find this evidence if certain types of microcredits have a significant difference in succes rate. If we find evidence supporting varying success rates among microcredit types, our second hypothesis seeks to identify key determinants contributing to microcredit success. Specifically, our research aims to determine if certain types of microcredits, distinguished by interest rates and borrower diversity, have significantly different success rates. Alternatively, we explore the possibility that no single factor consistently enhances microcredit performance.

The literature emphasizes the importance of innovative repayment structures, competitive interest rates, and borrower demographics in determining microcredit success (Mosley & Hulme, 1998; Banerjee, 2013). Successful microfinance programs often serve diverse borrower groups while maintaining low default rates (Morduch, 1999). On top of that, competitive interest rates, particularly below 30% per year (Banerjee, 2013), are associated with successful microcredit programs. By investigating these two success factors, we aim to provide actionable insights for program improvement and poverty alleviation efforts in Sub-Saharan Africa. This leads us to the second hypothesis and sub-hypotheses:

Hypothesis 2: *there are no certain characteristics that create a higher chance of a microcredit being more successful.*

Hypothesis 2 alternative: *there are certain characteristics that create a higher chance of a microcredit being more successful.*

To further specify this investigation, we propose two sub-hypotheses:

Hypothesis 2a: *microcredits with lower interest rates do not differ in success rates in comparison with microcredits with higher interest rates.*

Hypothesis 2a alternative: *microcredits with lower interest rates have a higher success rate than microcredits with a higher interest rates.*

Hypothesis 2b: *microfinance institutions with a diverse group of borrowers do not differ in success rates in comparison with microfinance institutions with a less diverse group of borrowers.*

Hypothesis 2b alternative: *microfinance institutions with a diverse group of borrowers have a higher success rate than microfinance institutions with a less diverse group of borrowers.*

These sub-hypotheses allow us to test specific factors within the broader framework of Hypothesis 2, providing a more detailed understanding of what influences microcredit success.

3. DATA & METHODS

3.1 Context

Microfinance in Sub-Saharan Africa is of great importance due to the region's significant poverty rates and limited access to formal financial services. With over 400 million people living below the international poverty line in Sub-Saharan Africa (World Bank, 2020), microfinance presents a promising avenue for financial inclusion and poverty alleviation. However, the effectiveness of microfinance interventions varies, with some models failing to adequately serve the poorest segments of society. Criticisms include concerns about profit-driven approaches that prioritize better-off clients and high-interest rates leading to over-indebtedness (Banerjee, 2013; Bateman, 2010).

Given these challenges, this paper focuses on researching microfinance in Sub-Saharan Africa to identify models with higher success rates, characterized by lower default rates and consequently meaningful poverty reduction impacts. By addressing these issues, this research aims to contribute to the understanding of effective microfinance strategies that can empower the poor and foster sustainable development in the region.

3.2 Data collection

The data for this research will be collected from the Microfinance Exchange Market (MIX) database. The MIX database is a globally accessible platform providing valuable microfinance information (World Bank Group, 2019). This database has information of publicly listed MFIs around the whole world for the years 2001-2019 and is the biggest dataset available in Sub-Saharan regarding microfinance. The initial dataset comprised comprehensive financial data for 69,367 observations of MFIs (including multiple observations of 60 to 90 variables per MFI), spanning the years 2015 to 2019. However, the year 2019 does not contain any data for the listed MFI's and thus cannot be used for the analysis. The year 2018 is therefore most recent and contains over 30 observations when considering the variables of interest. The years 2017, 2016 and 2015 contain each 11 observations or less considering the variables of interest. Therefore, for this research, since one of the objectives is to provide more recent information about microfinance, the year 2018 will be used for this research in Sub-Saharan Africa.

This dataset contains all the data needed regarding this research. This implies at least the value of the loan loss rate (%), to determine the success rate, and other variables that will be used to determine different types of microcredits, such as average loan balance per borrower, average number of loans outstanding, percent of female borrowers, cost per borrower, and different types of loans which are 1. Enterprise finance (1.1 Small & Medium sized 1.2 Microenterprise), 2. Household financing, 3. Location (3.1. Rural 3.2. Urban) and 4. Methodology, Individual.

One possible ethical problem with the MIX database is that many publicly listed MFI variables contain input *NULL* for the variables. After keeping all the data from the year 2018, removing unnecessary variables and *NULL* inputs, the dataset leads to having in total 31 observations with 22 descriptive variables. More variables will be removed in the process to make sure the model does not contain multicollinearity and more observations can be contained.

3.3 Empirical Model

To begin exploring the various types of microcredits, an exploratory cluster analysis will be conducted on the dataset. Utilizing the (Morissette & Chartier, 2013) k-means clustering algorithm, implemented in the kmeans function from the stats library in R, the data will be partitioned into k sets with the goal of minimizing within-cluster variance. The choice of the number of clusters, k, is determined by the user. To increase the probability of finding the optimal result, the number of random starts will be set to 25. Prior to clustering, the data will be normalized to address differences in levels and variances between variables, ensuring each variable has zero mean and unit variance. Subsequently, the clustering algorithm will be applied to the normalized data.

As a second step, we will analyze differences between clusters using a multinomial logistic regression model (Biemann et al., 2012; Ruef et al., 2003). This model examines the log odds of an observation belonging to a specific microcredit type (cluster) relative to a reference cluster based on a set of predictor variables. Specifically, the log odds of an observation being in a particular cluster rather than the reference cluster will be modeled based on predictor variables as follows:

$$y_i = \beta_0 + \beta_1 \times Lossrate_i + \cdots + \varepsilon_i$$

where y_i represents the log odds of belonging to a cluster other than the reference cluster. β_0 is the constant. $Lossrate_i$ is a continuous variable indicating the percentage of lost loans and ε_i is the error term of multinomial logistic regression analysis. The final model with the other variables will be determined after the correlation matrix and cluster analysis. Highly correlated variables will be removed, and the final multinomial logistic regression model will remain. The estimated parameters β_1 to β_x indicate how the likelihood of belonging to a different cluster change with a one-unit increase in the associated predictor variable. This indicates the direction, positive or negative, and the significance of the relationship.

Two-sided t-values will indicate whether the mean within a cluster significantly differs from the overall sample mean. An absolute t-value exceeding the threshold (e.g., at the 5% level) indicates a significant difference between the cluster mean and the overall mean.

3.4 Variables

In total 22 variables will be used in the descriptives analysis of microfinance institutions (MFIs) in Sub-Saharan Africa. The variables measure various dimensions of MFI performance, operational characteristics, and demographic characteristics, providing a comprehensive framework for the cluster analysis. All 22 variables will be discussed shortly. The loan loss rate (*Lossrate*) is a continuous variable indicating the percentage of loans written off as losses, which is the variable of most interest in this research. The different characteristics of the clusters with different success rates will be discussed and give us insights on what impact they possibly have on the success rates. Average loan per borrower (*AVGloan*) measures the average size of loans disbursed, influencing outreach and sustainability (Armendáriz & Morduch, 2006). The average number of borrowers (*AVGborrowers*) represents the average number of active borrowers, a key indicator of scale (Ledgerwood, 1999). Average number of loans outstanding (*AVGnumberloans*) indicates the average number of loans currently outstanding, reflecting lending practices (Rhyne & Otero, 2006). The ratio of average outstanding balance to GNI per capita (*ratio_balance/GNI*) measures the average outstanding loan balance relative to Gross National Income (GNI) per capita, providing insight into the debt burden on borrowers (Morduch, 1999). The ratio of gross loan portfolio to total assets (*ratio_GrLoan/TA*) represents the gross loan portfolio as a percentage of total assets, indicating lending emphasis (CGAP, 2003). Net income before taxes and donations (*NIBT&donations*) measures financial performance, while the percentage of female borrowers (*%femaleborrowers*) assesses gender outreach, which according to Pitt & Khandker (1998) has a more positive effect on social impact than loans for men. The profit margin (*profitmargin*) and retained earnings (*retainedearnings*) indicate financial sustainability and reinvestment capacity (Rosenberg, 2009; Ledgerwood & White, 2006). Risk coverage (*riskcov*) measures the loan portfolio's risk coverage (CGAP, 2003), and staff turnover rate (*staffturnover*) impacts operational efficiency

(Schreiner, 2002). Total Equity (*TotalEq*) indicates financial robustness (Ledgerwood, 1999), and Real Yield on Gross Portfolio (*realyield*) reflects actual portfolio returns and thus real interest rates on those loans (MicroRate, 2003).

Eight variables are added to the analysis to be able to categorize the institutions through type of loans and geographical distribution of the loans. The variables are created by dividing the original variables with real values by the variable 'average outstanding loan balance' and this gives a percentage of the part of the total amount of loans which belongs to a certain category. The percentage of enterprise finance overall (*%Enterprisefinance*), are divided in SME finance (*%SME*), and microenterprise finance (*%Microenterprise*) and they indicate the distribution of financing across different business types, crucial for economic development (Beck et al., 2007; Morduch, 1999). It is interesting to see whether a focus on microfinance or SME finance has a different correlation with success rates. The percentage of household finance (*%Householdfinance*) measures household lending, essential for meeting basic needs (Collins et al., 2009). The percentage of rural finance (*%Rural*) and urban finance (*%Urban*) capture the geographical distribution of loans, addressing the financial needs of rural and urban populations (Zeller & Meyer, 2002; Ledgerwood, 1999). The percentage of individual loans (*%Individual*) indicates the prevalence of individually tailored loans, enhancing financial inclusion (Morduch, 1999). These variables are interesting in measuring whether certain types of microloans have a correlation with a higher success rate.

These variables provide a robust framework for analyzing the differences between clusters of MFIs. By integrating these diverse variables, we can evaluate the performance, types of microcredits, and operational characteristics of MFIs in Sub-Saharan Africa, contributing to a deeper understanding of their impact and effectiveness. This approach aligns with methodologies used in prior research (Biemann et al., 2012; Ruef et al., 2003), ensuring the importance and relevance of the cluster analysis.

3.5 Descriptive statistics

The variables of interest for the models are stated on the next page. The mean, standard deviation, minimum and maximum can be seen per variable.

TABLE 1: DESCRIPTIVE STATISTICS

Variables	N	mean	sd	min	max
Loan Loss Rate (*Lossrate*)*	37**	1.962%	2.14%	-0.58%	7.16%
Average loan per borrower (*AVGloan*)	31	992.9	986.2	21.0	3291.0
Average number of active borrowers (*AVGborrowers*)	31	52,314	118,920	1177	660,043
Average number of loans outstanding (*AVGnumberloans*)	31	55,084	120,927	1277	660,043
Average outstanding balance / GNI per capita (%) (*ratio_balance/GNI*)	31	126.30%	173.16%	6.25%	680.78%
Cost per borrower (*costborrower*)	31	352.1	682.5	22.0	3820.0
Debt to equity ratio (*ratio_D/E*)	31	3.801	3.475	0.29	15.37
Gross loan portfolio to total assets (%) (*ratio_GrLoan/TA*)*	37**	73.37%	13.22%	48.10%	96.80%
Net income before taxes and donations (*NIBT&donations*)	31	356,625	3,452,996	-10,314,598	12,452,728
Percent of female borrowers (%) (*%femaleborrowers*)*	37**	56.59%	25.77%	0.00%	99.99%
Profit margin (%) (*profitmargin*)*	37**	3.90%	21.79%	-69.49%	39.46%
Retained earnings (*retainedearnings*)	31	2,246,598	12,035,260	-7,917,823	63,166,476
Risk coverage (%) (*riskcov*)	31	76.51%	50.84%	15.18%	286.14%
Staff turnover rate (%) (*staffturnover*)	31	12.78%	9.97%	1.02%	47.64%
Total equity (*TotalEq*)	31	9,573,894	14,246,551	301,229	63,166,476
Yield on gross portfolio (real) (%) (*realyield*)*	37**	31.63%	15.02%	13.73%	81.29%
Number of loans outstanding, Credit Products, Enterprise Finance (%) (*%Enterprisefinance*)	31	89.78%	24.7%	0.0%	115.36%
Number of loans outstanding, Credit Products, Enterprise Finance, loans to SME (%) (*%SME*)*	37**	8.24%	22.22%	0.0%	114.08%
Number of loans outstanding, Credit Products, Enterprise Finance, Microenterprise (%) (*%Microenterprise*)	31	84.57%	27.88%	0.0%	115.36%
Number of loans outstanding, Credit Products, Household finance (%) (*%Householdfinance*)	31	9.71%	22.87%	0.0%	106.37%
Number of loans outstanding, Location, Rural (%) (*%Rural*)	31	39.36%	40.17%	0.0%	110.15%
Number of loans outstanding, Location, Urban (%) (*%Urban*)	31	55.51%	41.36%	0.0%	115.36%
Number of loans outstanding, Methodology, Individual (%) (*%Individual*)*	37**	62.97%	43.66%	0.0%	115.36%

Note: The variable of most interest is *Lossrate;* the other variables are *AVGloan, AVGborrowers, AVGnumberloans, ratio_balance/GNI, costborrower, ratio_D/E, ratio_GrLoan/TA, NIBT&donations, %femaleborrowers, profitmargin, retainedearnings, riskcov, staffturnover, TotalEq, realyield, %Enterprisefinance, %SME, %Microenterprise, %Householdfinance, %Rural, %Urban* and *%Individual.* * *indicates the* *variables that exist in the final model, ** indicates the difference in observations in the final model due to having less NA observations when keeping less variables.*

22 variables are included in the descriptives statistics. However, as can be seen in the correlation matrix (Appendix B1), many variables have a correlation of 0.6 or higher. Therefore, the variables that will be kept are chosen due to theoretical significance and statistical importance. Loan loss rate (*Lossrate*) and the real yield on gross portfolio (*realyield*) should be kept due to the theoretical significance, other variables are kept due to statistical importance. The variables that have been chosen to use for the cluster analysis are loan loss rate (*Lossrate*), debt to equity ratio (*ratio_D/E*), gross loan portfolio to total assets (*ratio_GrLoan/TA*), percent of female borrowers (*%femaleborrowers*), profit margin (*profitmargin*), staff turnover rate (*staffturnover*), real yield on gross portfolio (*realyield*), number of loans outstanding to SME (*%SME*) and number of individual loans (*%Individual*).

The loan loss rate has an average of 1.962%, with a minimum of -0.058% and a maximum of 7.16%. This indicates that on average 1.962% of the loans are not repaid. Notably, one observation reflects a negative loan loss rate of -0.058%, suggesting a repayment exceeding the loan amount by 0.058%. This negative percentage can be explained by how the loan loss rate is measured, namely: (Write offs – Value of Loans Recovered) / Average Gross Loan Portfolio (World Bank Group, 2019). The value of loans recovered is apparently higher than the write offs in that period for one observation. Since this could be a plausible explanation, it is chosen to keep this observation. The international benchmark for loan loss rates stands at 3% (Addae-Korankye, 2014). Notably, in Ghana, approximately 60% of Microfinance Institutions (MFIs) exhibit loan loss rates surpassing this threshold (Addae-Korankye, 2014). Therefore, the average loan loss rate of 1.962% in this sample might be lower than the general population average which should be considered when analyzing the results. On top of that, according to Morduch (1999), successful MFIs maintain low default rates, often indicated as below 10% and in some cases below 2%. Since no observation has a default rate higher than 7.16%, the dataset contains already quite successful microloans.

The real yield on the gross portfolio has a mean of 31.63% and varies from 13.73% to 81.29%. Interest rates in microcredit programs exhibit significant regional variations. For instance, while regions such as Mexico and South Africa experience interest rates exceeding 100% annually, successful microfinance models in countries like Bangladesh, Bolivia, India, and Indonesia typically maintain interest rates below 30% per year (Banerjee, 2013). Therefore, in this sample, it seems in line with the population mean according to literature.

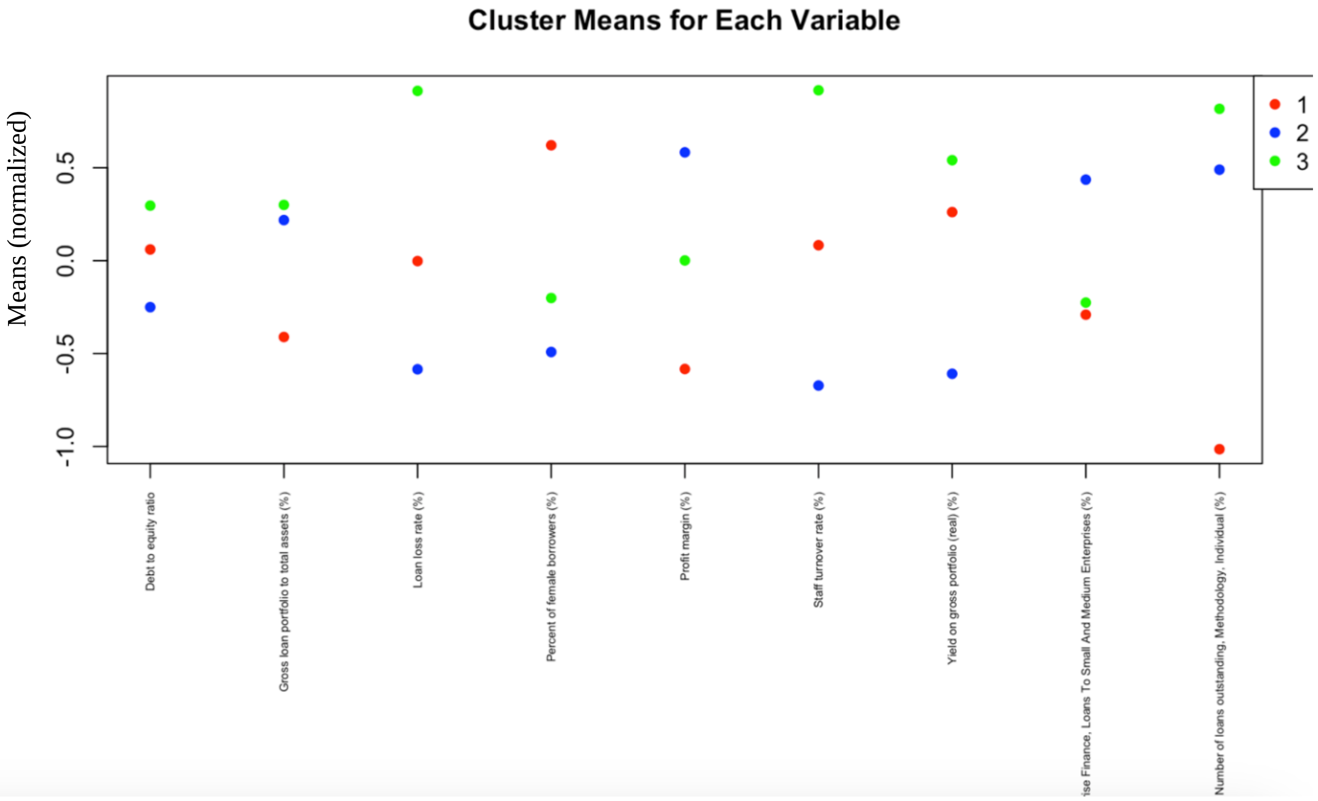
The variables of SME loans (*%SME*) and Individual loans (*%Individual*) both contain maximum observations of above 100%. This could be explained by the creation process of these variables: dividing the ‘number of loans by type’ by the ‘average outstanding loan balance’. These variables then give a relative percentage of the loans by this type, instead of an absolute number. However, since it is divided by the ‘average outstanding loan balance’, it could be that the average of the year, is lower than the outstanding balance of these types of loans at a certain period in time. Therefore, the observed values of *%SME* and *%Individual* in this sample could be higher than the actual values in the population.

In summary, the descriptive statistics for the variables within this sample do not exhibit significant deviations from established population means. Therefore, these variables are deemed reliable for further analysis and interpretation of results.

4. RESULTS & INTERPRETATION

This section provides the results of the cluster analysis and the multinomial logistic regression analysis. First, figure 1 displays the results of the first cluster analysis. One of two variables that are very alike, will then be chosen to be kept to continue with a better fitting model. Secondly, table 2 displays the results of the applied multinomial logistic regression analysis including the normalized means for each cluster and the coefficients for cluster 2, and 3. Thirdly, the model outcomes will be explained and a robustness analysis will be done. And lastly, we discuss the model outcomes extensively and connect these to earlier literature research.

FIGURE 1: CLUSTER PROFILES RESULTS



Note: number 1, 2, and 3 represent respectively cluster 1, 2, and 3. The y-axis shows the normalized means of the results. After this analysis, the variables ‘Staff turnover rate’ and ‘Debt to equity ratio’ are chosen to be removed.

Figure 1 displays cluster 1, 2 and 3 with 9 variables that have stayed after the correlation analysis. Variables that lookalike in the distribution between cluster 1, 2 and 3 will be removed due to showing similar cluster means. Loan loss rate (*Lossrate*) and staff turnover rate (*staffturnover*) have almost identical results for all clusters. It is chosen to remove staff turnover rate (*staffturnover*) for the multinomial logistic regression, since Loan loss rate (*Lossrate*) is a variable of greater interest in this research. The same similarities apply to the variables debt to equity ratio (*ratio_D/E*) and real yield to gross portfolio (*realyield*). Since the variable *realyield* can be linked to interest rates, in which it is interesting to measure the relationship between success rates and interest rates, which is negative according to (Banerjee, 2013), it is chosen to keep *realyield* for the multinomial regression analysis due to theoretical importance. The multinomial logistic regression analysis will be continued with the remaining 7 variables.

Interesting patterns can be seen in cluster 1, 2 and 3. Cluster 1 displays the most average loan loss rate (*Lossrate*) of the 3, with a relatively high number of female borrowers (*%femaleborrowers*), contains the lowest profit margin (*profitmargin*) and the least focus on individual loans (*%individual*). We label this cluster as the 'moderate performers', since they have a moderate loan loss rate.

Cluster 2 has the lowest loan loss rates (*Lossrate*), which is equal to the highest success rate in this study, has the highest profit margin (*profitmargin*), but at the sime time the lowest real yield on the gross portfolio (*realyield*). On top of that, cluster 2 has the highest focus on SME (*%SME*) loans in comparison with the other clusters. We label this cluster as the 'high performers', since they have a low loan loss rate.

Cluster 3 displays the higest loan loss rate (*Lossrate*), which is in line with being the least successful cluster. It also has the most focus on individuals (*%individual*) of the 3 clusters, and has the highest real yield on gross portfolio (*realyield*). We label this cluster as the 'low performers', since they have a high loan loss rate.

TABLE 2: CLUSTER RESULTS, MULTINOMIAL LOGISTIC REGRESSION

	1:	**2:**		**3:**	
	Cluster mean (normalized)	Cluster mean (normalized)	Coefficient (b’s)	Cluster mean (normalized)	Coefficient (b’s)
Intercept			-8.1830 (0.9993)		-5.0488 (0.9999)
Loan Loss Rate (*Lossrate*)	-0.0024 (0.9940)	-0.5848 (0.0001)***	-18.2001 (0.9996)	0.9133 (0.0051)***	6.3039 (0.9998)
Gross loan portfolio to total assets (%) (*ratio_GrLoan/TA)*	-0.4111 (0.0795)*	0.2183 (0.5178)	8.5182 (0.9970)	0.2999 (0.2412)	5.5915 (0.9998)
Percent of female borrowers (%) (*%femaleborrowers*)	0.6208 (0.0061)***	-0.4915 (0.1479)	-16.3940 (0.9967)	-0.2012 (0.2298)	-16.7913 (0.9993)
Profit margin (%) (*profitmargin*)	-0.5833 (0.0608)*	0.5827 (0.0024)***	12.8207 (0.9901)	0.0008 (0.9978)	6.7324 (0.9999)
Yield on gross portfolio (real) (%) (*realyield*)	0.2611 (0.4399)	-0.6087 (0.0002)***	-36.0152 (0.9941)	0.5407 (0.0983)*	7.7654 (0.9992)
Number of loans outstanding, Credit Products, Enterprise Finance, loans to SME (%) (*%SME*)	-0.2909 (0.0001)***	0.4360 (0.3104)	3.4349 (0.9999)	-0.2258 (0.0003)***	-0.8073 (0.9999)
Number of loans outstanding, Methodology, Individual (%) (*%Individual*)	-1.0148 (0.000)***	0.4896 (0.0423)**	20.9305 (0.9969)	0.8170 (0.001)***	21.9566 (0.9982)
#Observations	14	14		9	

Note: The column labelled cluster mean shows the estimated mean of a variable within the cluster. Between the parentheses is the t-value of the difference between the cluster mean the full sample mean (i.e. over all clusters). The coefficients b's show how the log odds of being in cluster 2 or 3 instead of being in the reference cluster 1 changes when the variable increases by one unit. The parentheses *p <0,10; **p<0,05; ***p<0,01 indicate the significance levels of the regression standard errors of respectively 10, 5 and 1 per cent. The variable of most interest is *Lossrate;* the other variables are *ratio_GrLoan/TA*, *%femaleborrowers, profitmargin, realyield, %SME* and *%Individual.*

4.1 Results

The outcomes of the multinomial logistic regression, coupled with cluster means can be seen in table 2. The selection of the number of clusters was guided by the Silhouette method (Appendix C2). Cluster 1, 2 and 3 have respectively 14, 14 and 9 observations, which is relatively similar and is in line with literature that states that a similar number of observations per cluster is preferred (Malinen & Fränti, 2014).

To examine the differences between these three clusters more objectively than described in the cluster analysis above, we test the difference between these microloans in two ways: (1) we test whether the differences of mean normalized values of variables are significantly different from the whole sample and (2) the differences of clusters 1, 2 and 3 are analyzed with a multinomial logistic regression model.

In Table 2, we first describe the characteristics of cluster 1 (‘moderate performers’). The cluster means show us that the variable *%femaleborrowers* is significantly higher (at the 1%

significance level) in cluster 1 than in the overall sample. The following variables have a significantly lower mean in cluster 1 than overall: the percentage of loans in SME (*%SME*) and invidual loans (*%individual*). At the 10% significance level, cluster 1 has a lower normalized mean on gross loan portfolio to total assets (*ratio_GrLoan/TA*) and a lower profit margin (*profitmargin*).

Cluster 2 ('high performers') is significantly higher than the overall sample in profit margin (*profitmargin*) and loans for individuals (*%individuals*). Besides, the means are significantly lower for the variables loan loss rate (*Lossrate*) and the real yield on gross portfolio (*realyield*).

Cluster 3 ('low performers') show a significantly high mean for loan loss rate (*Lossrate*), individual loans (*%individual*) and (at the 10% level) for the real yield on gross portfolio (*realyield*), while significantly scoring lower on SME loans (*%SME*).

Hypothesis 1: '*types of microcredits do not differ in success rates',* can be rejected when looking at the differences in cluster means of *Lossrate* of cluster 2 and cluster 3. The two-sided t-value test show significant differences in cluster 2 and 3 in *Lossrate* at the 1% significance level.

However, when looking at the multinomial logistic regression analysis, no result is deemed significant and hypothesis 2: *there are no certain characteristics that create a higher chance of a microcredit being more successful,* can thus not be rejected.

In summary, cluster 1 shows the 'moderate performers' with the characteristics of having a high focus on female borrowers and a significantly lower focus on SME loans and individual loans. Cluster 2, the 'high performers', have the characteristics of having a high profit margin and a focus on individual loans. However, the real yield on gross portfolio is significantly lower. And lasty, cluster 3 ('low performers') has a high focus on individual loans, and scores lower on SME loans.

4.2 Robustness check

A robustness check has been done by doing the same cluster analysis and multinomial logistic regression analysis with the same variables (Appendix E1). These analysis are chosen for the year 2017 and will be compared with the analysis of the year 2018. 2017 is chosen since it had the most observations in comparison with the years 2016 and 2015, namely 11 observations in comparison with 9 and 8 observations.

Cluster 1, 2, and 3 have respectively 1, 6 and 4 observations. Therefore cluster 1 is treated as an outlier. Due to the little number of observations and non-signficance, it is impossible to draw conclusions from the robustness analysis. However, cluster 2 of the robustness analysis looks most like cluster 3 of the 2018 analysis, with a (non-significant) positive *Lossrate*. And cluster 3 of the robustness check is most similar with cluster 3 of the 2018 analysis, the 'high performers'. The 'low performers' of year 2017 and 2018 seem similar in the lower focus on SME loans and having a higher profit margin, however there focus on individual loans differs. For the 'high performers' comparing 2017 and 2018, they seem only similar in the positive profit margin, but differ in the real yield and focus in individuality.

In summary, no conclusions can be drawn from the robustness of the model, since too little observations and no significance is present. This forms a limitation in the results, since the robustness cannot be measured.

4.3 Interpretation

The results from the analysis provide a nuanced understanding of the different characteristics and success factors of microcredit institutions, particularly in Sub-Saharan Africa. This section interprets these findings in line with existing literature.

First, it is important to mention that only results from the cluster analysis are statistically significant, no conclusions can thus be drawn from the multinomial logistic regression analysis. This aligns with findings from recent studies that emphasize the contextual variability and complex operational dynamics within microfinance institutions (Hermes & Lensink, 2011; Cull et al., 2009). The insignificance of these results is consistent with the challenges of achieving statistically significant outcomes in heterogeneous microfinance environments, as discussed by Armendáriz and Morduch (2006) and Karlan and Zinman (2011).

The significant differences in loan loss rates between cluster 2 and 3 reject hypothesis 1: '*types of microcredits do not differ in success rates',* since significant differences are present in *Lossrate* between the clusters. This finding aligns with the literature, which highlights the diverse performance outcomes of microcredit institutions based on varying operational models and borrower demographics. Studies have shown that tailored financial products and innovative repayment structures significantly impact microfinance success (Matin et al., 2002; Hulme and Mosley, 1996).

Cluster 1 is characterized by lower percentages of SME and individual loans, suggesting a more diversified or conservative loan portfolio. The cluster means reveal a higher percentage of female borrowers, which aligns with previous findings where a higher proportion of female borrowers can indicate a focus on empowerment but may also correlate with higher risk due to the socio-economic vulnerabilities often associated with female borrowers (Armendariz & Morduch, 2010). Furthermore, the negative profit margin indicates that the institution is not profitable, highlighting the financial struggles and low profitability of these microfinance institutions. This cluster aligns with the second wave of microfinance, where loans were targeted towards women running microenterprises, but highlights the associated risks and lower profitability in certain contexts (Matin et al., 2002; Hulme & Mosley, 1996).

Cluster 2, the 'high performers', exhibits higher profitability and a significant focus on individual loans, which can be associated with higher risk but also higher profitability if managed well. The mean values show that Cluster 2 significantly outperforms the overall sample in terms of profit margin and individual loans. The significantly lower real yield on the gross portfolio, although counterintuitive, may reflect lower interest rates that align with sustainable financial practices (Banerjee, 2013). This is in line with literature of Banerjee (2013), suggesting that competitive interest rates, are associated with successful microcredit programs. The high profitability and low loss rates corroborate the literature on effective microcredit management (Mosley & Hulme, 1998). This cluster's emphasis on innovative repayment structures and competitive interest rates reflects the successful models in countries like Bangladesh and India, where interest rates are maintained below 30% per year, contributing to the sustainability and success of microfinance programs (Banerjee, 2013; Morduch, 1999).

Cluster 3, the 'low performers', shows a high loan loss rate and a significant focus on individual loans, reflecting higher financial risk and lower success. On top of that, it shows a significant lower focus on SME loans. The cluster's characteristics include high loan loss rates and real yields on gross portfolios, indicating mispricing or high-risk lending strategies. The high loan loss rates and higher interest rates (due to the higher *realyield*) align with Banerjee (2013), which associates higher interest rates with higher loan loss rates and thus lower success. This cluster reflects the first wave of microfinance, where subsidized credit aimed at small farmers was believed to increase productivity and incomes but often resulted in high default rates due to the inability to repay (Matin et al., 2002).

In conclusion, the new results indicate clear differences among the cluster means. Cluster 1 ('moderate performers') demonstrates balanced performance with moderate risks, reflecting the mixed outcomes of targeting female borrowers and diversification strategies. Cluster 2 ('high performers') exhibits high profitability and lower interest rates, reflecting effective management and pricing strategies aligned with the best practices highlighted in the literature. Cluster 3 ('low performers') shows high loss rates, indicating potential mismanagement or high-risk portfolios. All findings are in line with Banerjee (2013), suggesting that more successful microloans have lower interest rates on the loans. Hypothesis 1 which tests differences in success rates between microloans can be rejected. However, hypothesis 2, which states that no certain characteristics that create a higher chance of a microcredit being more successful, cannot be rejected due to insignificant multinomial logistic regression results.

5. DISCUSSION & CONCLUSION

This study aimed to investigate the various types of microcredits implemented in Sub-Saharan Africa and their respective success rates, guided by the central research question: "*Which types of microcredits are implemented in Sub-Saharan Africa, and how do these types of microcredits differ in terms of success?*" Through an analysis of key characteristics and success factors of microcredit institutions, this study provides insights for microfinance institutions and policymakers alike.

5.1 Discussion of findings

The findings from this study reveal significant insights into the performance and characteristics of microcredit institutions across different clusters. Cluster analysis identified three distinct groups: moderate performers (cluster 1), high performers (cluster 2), and low performers (cluster 3).

Cluster 1 represents institutions characterized by moderate performance metrics. These institutions exhibit a balanced approach with moderate loan loss rates and a conservative lending strategy. They notably have a higher proportion of female borrowers compared to the overall sample, indicating a focus on gender empowerment initiatives. However, they show lower concentrations in SME loans and individual loans relative to the overall sample. The negative profit margin suggests financial challenges.

Cluster 2, identified as high performers, demonstrates lower loan loss rates and higher profitability. These type of microloans have a significant focus on invidual loans. The lower real yield on gross portfolio is in line with Banerjee's (2013) study in which more successful microcredits have competitive interest rates.

Cluster 3 represents institutions categorized as low performers, characterized by higher loan loss rates and a focus on individual loans. These institutions have a significant focus on individual loans and a lower emphasis on SME loans suggesting a narrower loan portfolio, potentially limiting diversification benefits. Besides, the *realyield* of this cluster is significantly higher at the 10% significance level. This is in line with Banerjee (2013) that less successful microloans have higher interest rates.

5.2 Implications for microfinance institutions and policymakers

The findings imply several considerations for microfinance institutions and policymakers. Institutions should consider the different types of microcredits to have an impact on their success rate. This study shows that the success rates between different microloans are indeed significantly different from each other. No significant results have been found on the characteristics leading to the higher success rates of these microloans. However, all results are in line with Banerjee (2013), suggesting that competitive interest rates have a correlation with more successful microloans. Microfinance and policymakers should take this finding into account when aiming for successful microloans.

On top of that, the 'moderate' and 'low performers' both have a signficant, negative focus on SME loans, while the 'high performers' have a moderate focus on SME loans. These results suggest that SME loans in the portfolio might be important for successful microloans. However, there is no significant result for this possible correlation.

Microfinance insitutions and policy makers should carefully reflect on the characteristics of their microloans and adjust accordingly to have a higher success rate. Emphasizing long-term sustainability over immediate profitability can help institutions navigate high-risk environments and ensure their longevity, but also those of the borrowers for a higher and positive social impact.

5.3 Limitations and recommendations for future research

Several limitations must be acknowledged. The study's sample size is relatively small with a number of 37 observations, limiting the generalizability of findings, especially present in the multinomial logistic regression analysis. Future research should aim to replicate these findings with larger and more diverse datasets to improve reliability and robustness, for example not only looking at Sub-Saharan Africa, but add more continents to have a higher number of observations.

Another limitation is the research year 2018. Only the year 2018 has been researched in this study, since the data of year 2019 contained zero observations and no data was collected after that year. Eventhough 2018 is relatively recent, it still has to be aknowledged that the data is from more than half a decade ago.

Thirdly, the robustness check, could only be performed with 11 observations as a maximum in the year 2017, since the years 2016 and 2015 contained even less observations. Due to the low number of observations and non-significant results, the robustness check is not deemed reliable and thus forms a limitation of this research.

It is advised for future research to further explore additional factors influencing the success rates of microcredits. Potential areas for further investigation include examining whether different types of microcredits in enterprise finance (including SME and microenterprise), geographical differences (urban vs. rural) and a focus on individual loans have a significant impact on success rates. Conducting studies with larger sample sizes could improve the generalizability and statistical significance of these findings. Qualitative approaches could also be interesting to provide deeper insights into the contextual factors influencing microfinance institutions and their success rates.

5.4 Conclusion

This study contributes to the ongoing academic literature on microcredit by providing empirical insights into the implementation and effectiveness of microcredits in Sub-Saharan Africa. The findings highlight meaningful distinctions among microcredit institutions based on their operational models and loan portfolios. By addressing these aspects, microfinance institutions can better fulfill their potential as tools for poverty alleviation and sustainable development in Sub-Saharan Africa.

6. REFERENCES

Addae-Korankye A. (2014). Causes and Control of Loan Default/Delinquency in Microfinance Institutions in Ghana. *American International Journal of Contemporary Research, 4*(12)

Armendáriz de Aghion, B., & Morduch, J. (2006). The economics of microfinance. *MIT Press*.

Banerjee, A., Duflo, E., Glennerster, R., & Kinnan, C. (2015). The Miracle of Microfinance? Evidence from a Randomized Evaluation. *American Economic Journal: Applied Economics*, *7*(1), 22–53. http://www.jstor.org/stable/43189512

Banerjee, A. (2013). Microcredit under the microscope: What have we learned in the past two decades, and what do we need to know? *Annual Review of Economics*, *5*(1), 487–519. https://doi.org/10.1146/annurev-economics-082912-110220

Bateman, M. (2010). Why doesn't microfinance work? The destructive rise of local neoliberalism. *Zed Books*.

Beck, T., Demirguc-Kunt, A., & Levine, R. (2007). *Finance, Inequality, and the Poor*. Journal of Economic Growth, 12(1), 27-49. https://hdl.handle.net/10411/12885

Biemann, T., Zacher, H., & Feldman, D. C. (2012). Career patterns: A twenty-year panel study. *Journal of Vocational Behavior*, *81*(2), 159–170. https://doi.org/10.1016/j.jvb.2012.06.003

CGAP. (2003). *Definitions of Selected Financial Terms, Ratios, and Adjustments for Microfinance*. CGAP, The World Bank Group.

Collins, D., Morduch, J., Rutherford, S., & Ruthven, O. (2009). *Portfolios of the Poor: How the World's Poor Live on $2 a Day*. Princeton University Press.

Cull, R., Demirgüç-Kunt, A., & Morduch, J. (2009). Microfinance Meets the Market. Journal of Economic Perspectives, 23(1), 167-192.

Daley-Harris, S. (Ed.). (2009). State of the microcredit summit campaign report. *Microcredit Summit Campaign*.

Duflo, E. (2012). Women's empowerment and economic development. Journal of Economic Literature, 50(4), 1051-1079. https://doi.org/10.1257/jel.50.4.1051

Hermes, N., & Lensink, R. (2011). Microfinance: Its Impact, Outreach, and Sustainability. World Development, 39(6), 875-881.

Hulme, D., & Mosley, P. (1996). Finance Against Poverty: Volumes 1 and 2. *Routledge*. https://www.escholar.manchester.ac.uk/uk-ac-man-scw:4b42

Karlan, D., & Zinman, J. (2011). Microcredit in Theory and Practice: Using Randomized Credit Scoring for Impact Evaluation. Science, 332(6035), 1278-1284.

Mosley, P., & Hulme, D. (1998). Microenterprise finance: Is there a conflict between growth and poverty alleviation? *World Development, 26(5),* 783–790. https://doi.org/10.1016/s0305-750x(98)00021-7

Johnson, S., & Rogaly, B. (1997). *Microfinance and poverty reduction*. https://doi.org/10.3362/9780855988005

Ledgerwood, J. (1999). *Microfinance Handbook: An Institutional and Financial Perspective*. The World Bank.

Ledgerwood, J., & White, V. (2006). *Transforming Microfinance Institutions: Providing Full Financial Services to the Poor*. The World Bank.

Malinen, M. I., & Fränti, P. (2014). Balanced k-means for clustering. *In A. Fred, J. Filipe, & H. Gamboa (Eds.), Proceedings of the Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR) (pp. 32-41)*. Springer. https://doi.org/10.1007/978-3-662-44415-3_4

Matin, I., Hulme, D., & Rutherford, S. (2002). Finance for the poor: from microcredit to microfinancial services. *Journal of International Development*, *14*(2), 273–294. https://doi.org/10.1002/jid.874

Mersland, R., & Strøm, R. Ø. (2009). Performance and governance in microfinance institutions. Journal of Banking & Finance, 33(4), 662-669. https://doi.org/10.1016/j.jbankfin.2008.11.009

MicroRate. (2003). *Performance Indicators for Microfinance Institutions: Technical Guide*. Inter-American Development Bank.

Morduch, J. (1999). The microfinance promise. *Journal of Economic Literature, 37*(4), 1569–1614. https://doi.org/10.1257/jel.37.4.1569

Morissette, L., & Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology*, *9*(1), 15–24. https://doi.org/10.20982/tqmp.09.1.p015

Mosley, P., & Hulme, D. (1998). Microenterprise finance: Is there a conflict between growth and poverty alleviation? *World Development, 26*(5), 783–790. https://doi.org/10.1016/s0305-750x(98)00021-7

Pitt, M. M., & Khandker, S. R. (1998). *The Impact of Group-Based Credit Programs on Poor Households in Bangladesh: Does the Gender of Participants Matter?*. Journal of Political Economy, 106(5), 958-996. https://www.jstor.org/stable/10.1086/250037

Rhyne, E., & Otero, M. (2006). *Microfinance Through the Next Decade: Visioning the Who, What, Where, When and How*. ACCION International.

Rosenberg, R. (2009). *Measuring Results of Microfinance Institutions: Minimum Indicators That Donors and Investors Should Track*. CGAP.

Ruef, M., Aldrich, H. E., & Carter, N. M. (2003). The Structure of Founding Teams: Homophily, Strong Ties, and Isolation among U.S. Entrepreneurs. *American Sociological Review, 68*(2), 195–222. https://doi.org/10.2307/1519766

Schreiner, M. (2002). *Aspects of Outreach: A Framework for Discussion of the Social Benefits of Microfinance*. Journal of International Development, 14(5), 591-603.

United Nations. (2015). Transforming our world: The 2030 agenda for sustainable development. United Nations.

World Bank Group. (2019). *MIX Market | DataBank*. Retrieved March 24, 2024, from https://databank.worldbank.org/source/mix-market

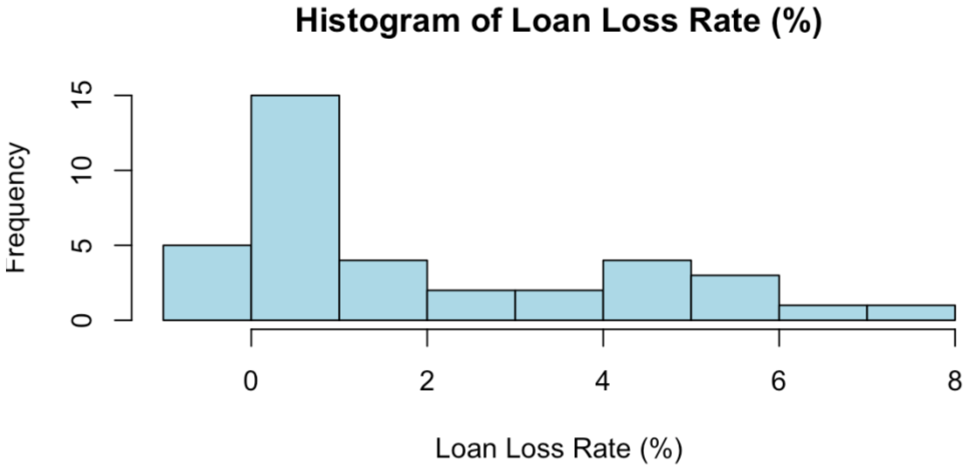
World Bank. (2020). Poverty and Shared Prosperity 2020: Reversals of Fortune. Washington, DC: World Bank.

Zeller, M., & Meyer, R. L. (2002). *The Triangle of Microfinance: Financial Sustainability, Outreach, and Impact*. Johns Hopkins University Press.

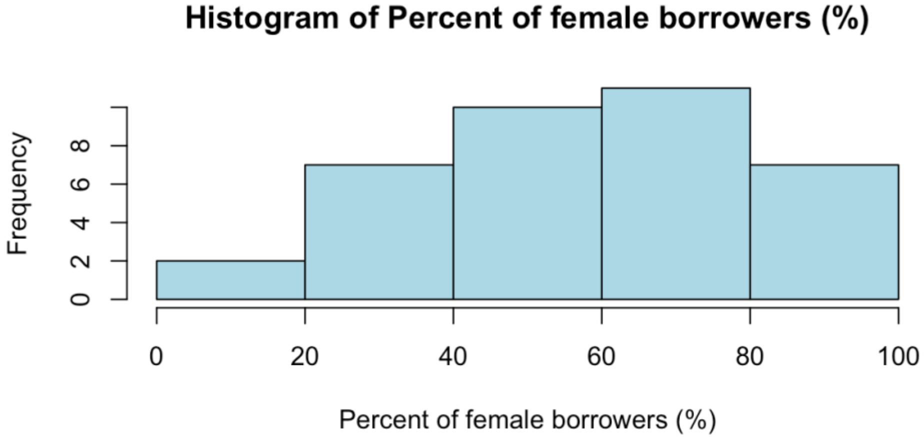
7. APPENDICES

APPENDIX A: DESCRIPTIVE STATISTICS

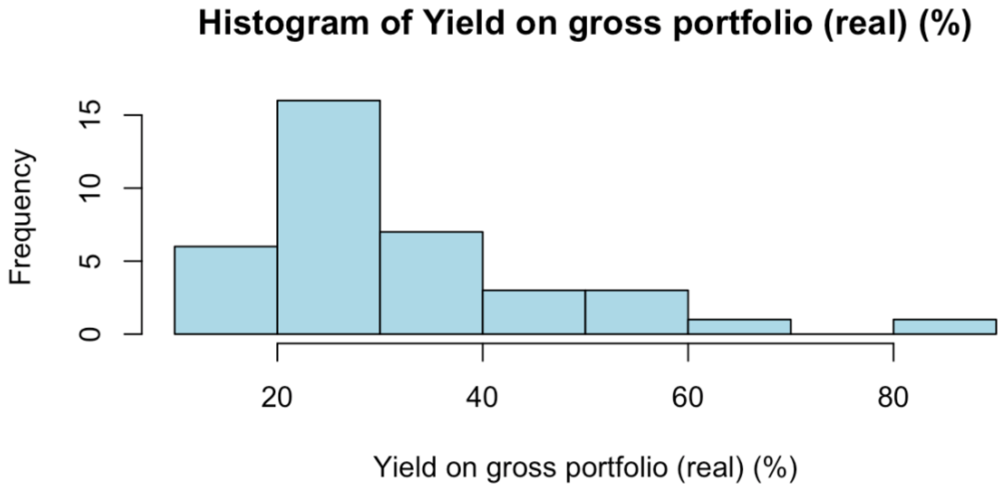
A1: Distribution of Loan Loss Rate (*Lossrate*)



A2: Distribution of percentage of female borrowers (*%femaleborrowers):*

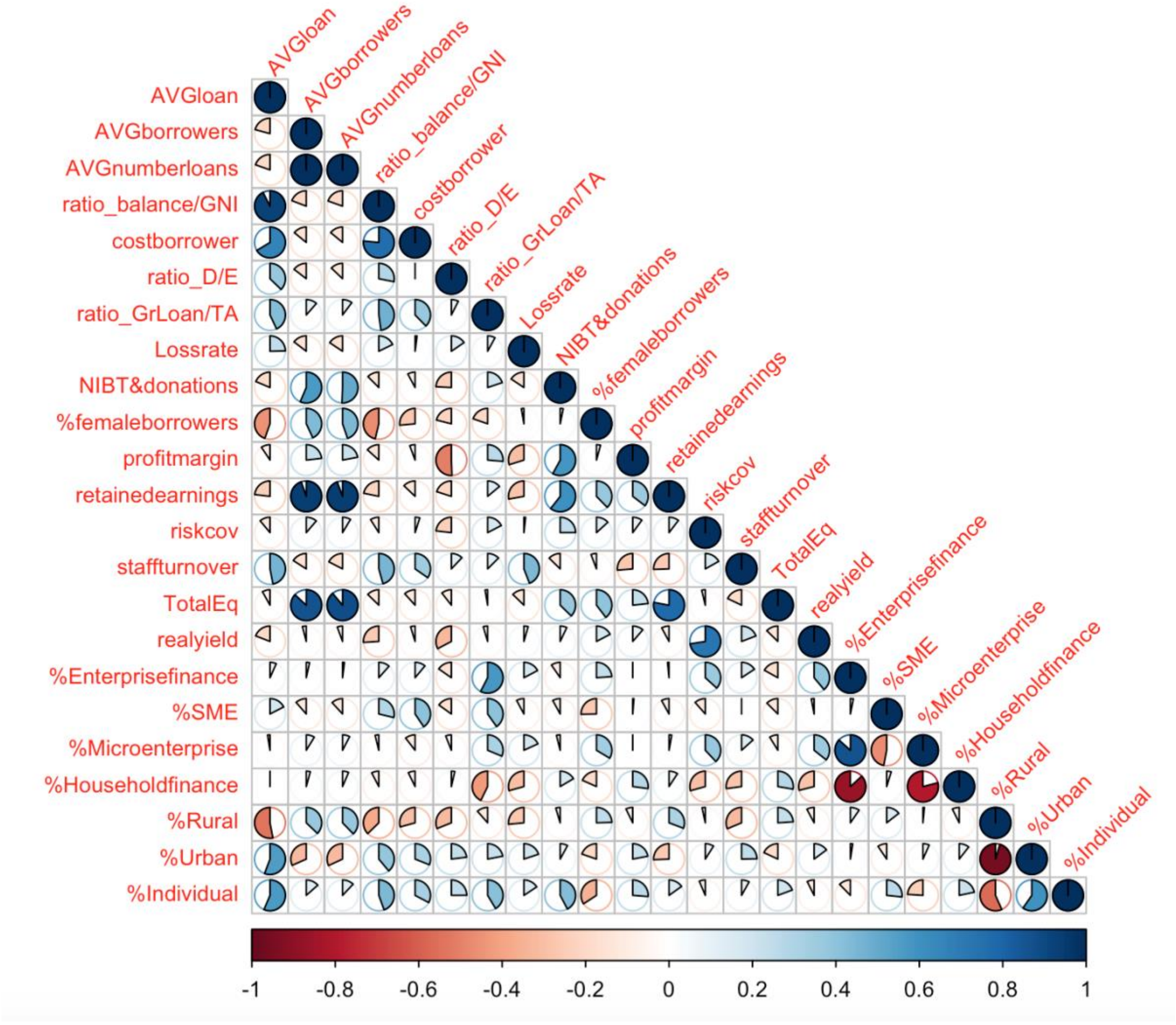


A3: Distribution of yield on gross portfolio (real) (*%realprofit*):



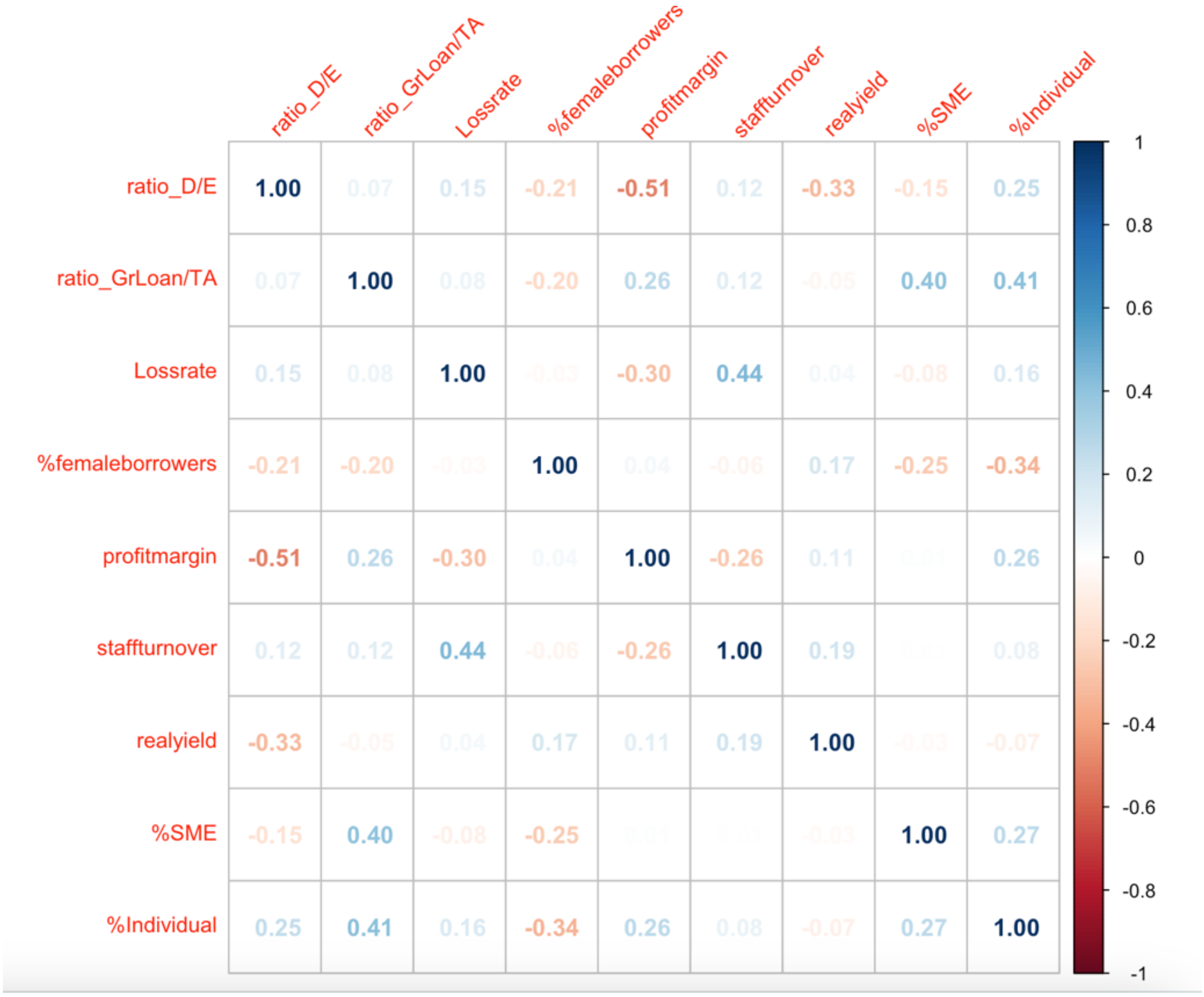
APPENDIX B: CORRELATION MATRIX

B1: Correlation Matrix before removement



Note: every variable that has a correlation of 0.6 or higher with another variable, will be elimated (1 of 2 will be kept). Chosen to be removed are: AVGloan, AVGborrowers, AVGnumberloans, ratio_balance/GNI, costborrower, retainedearnings, riskcov, TotalEq, %Enterprisefinance, %Microenterprise, %Householdfinance, %Rural and %Urban

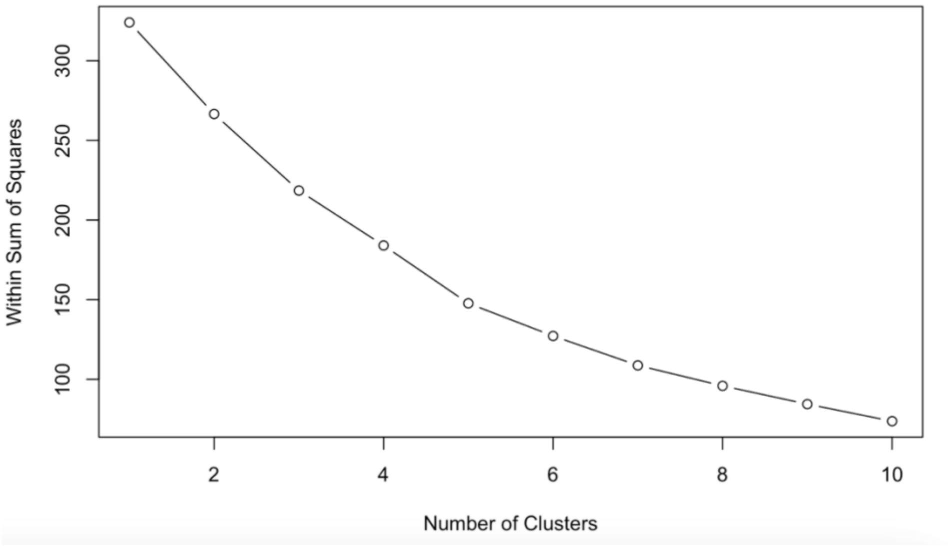
B2: Correlation Matrix after removement of high correlated variables



	ratio_D/E	ratio_GrLoan/TA	Lossrate	%femaleborrowers	profitmargin	staffturnover	realyield	%SME	%Individual
ratio_D/E	1.00	0.07	0.15	-0.21	-0.51	0.12	-0.33	-0.15	0.25
ratio_GrLoan/TA	0.07	1.00	0.08	-0.20	0.26	0.12	-0.05	0.40	0.41
Lossrate	0.15	0.08	1.00	-0.03	-0.30	0.44	0.04	-0.08	0.16
%femaleborrowers	-0.21	-0.20	-0.03	1.00	0.04	-0.06	0.17	-0.25	-0.34
profitmargin	-0.51	0.26	-0.30	0.04	1.00	-0.26	0.11	0.01	0.26
staffturnover	0.12	0.12	0.44	-0.06	-0.26	1.00	0.19	0.01	0.08
realyield	-0.33	-0.05	0.04	0.17	0.11	0.19	1.00	-0.03	-0.07
%SME	-0.15	0.40	-0.08	-0.25	0.01	0.01	-0.03	1.00	0.27
%Individual	0.25	0.41	0.16	-0.34	0.26	0.08	-0.07	0.27	1.00

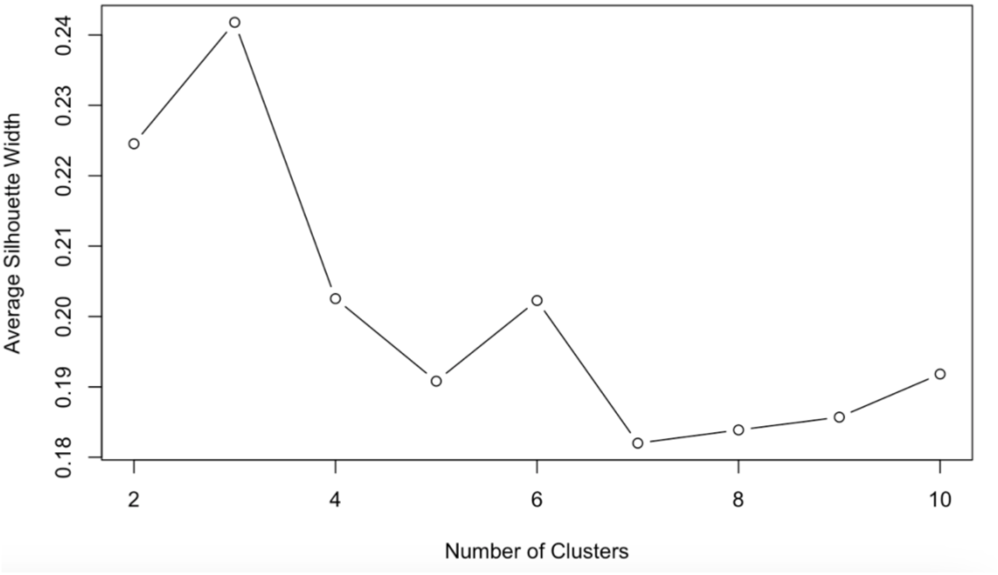
APPENDIX C: ELBOW & SILHOUETTE

C1: Elbow method



Note: since there is no clear 'elbow' in appendix A1, there has been chosen to do the Silhouette method in appendix A2.

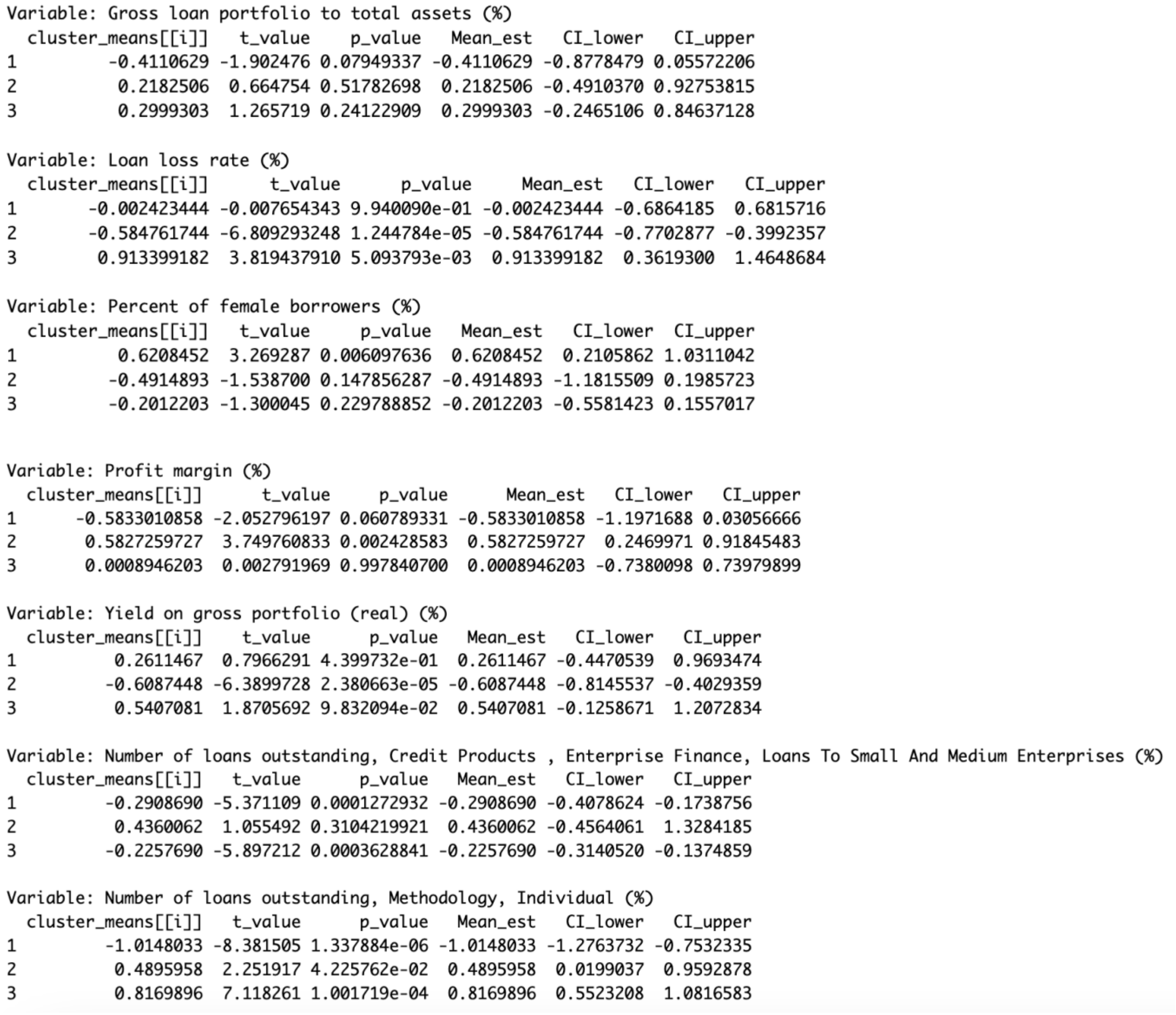
C2: Silhouette Scores



Note: the number of clusters with the highest Average Silhouette Width should be chosen as number of clusters, which is 4 in this case.

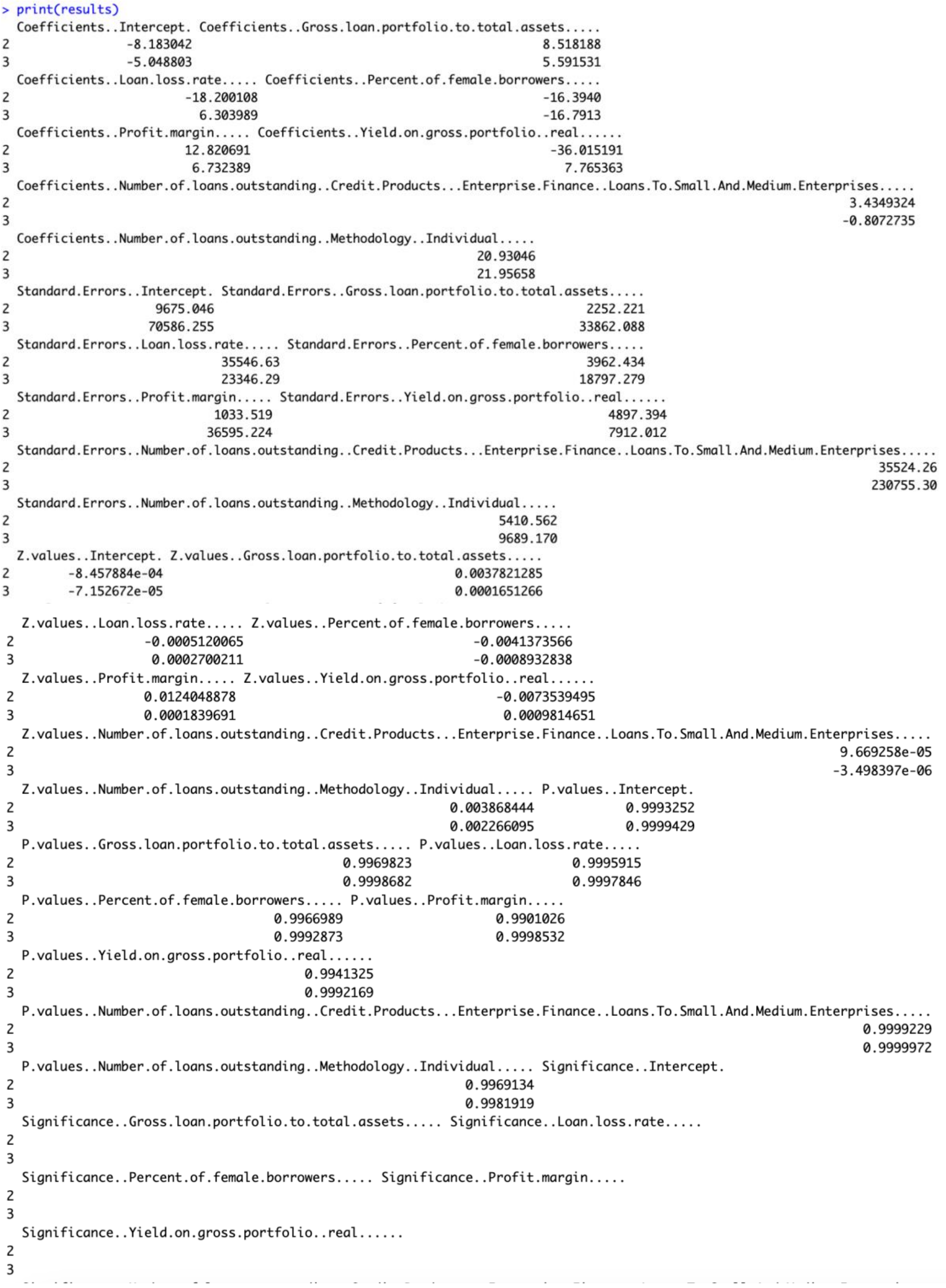
Appendix D: RESULTS

D1: Normalized cluster means (via R-studio)

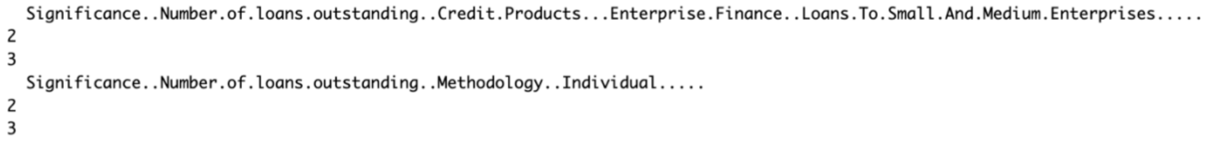


```
Variable: Gross loan portfolio to total assets (%)
  cluster_means[[i]]   t_value    p_value   Mean_est   CI_lower   CI_upper
1         -0.4110629 -1.902476 0.07949337 -0.4110629 -0.8778479 0.05572206
2          0.2182506  0.664754 0.51782698  0.2182506 -0.4910370 0.92753815
3          0.2999303  1.265719 0.24122909  0.2999303 -0.2465106 0.84637128

Variable: Loan loss rate (%)
  cluster_means[[i]]      t_value      p_value     Mean_est   CI_lower   CI_upper
1       -0.002423444 -0.007654343 9.940090e-01 -0.002423444 -0.6864185  0.6815716
2       -0.584761744 -6.809293248 1.244784e-05 -0.584761744 -0.7702877 -0.3992357
3        0.913399182  3.819437910 5.093793e-03  0.913399182  0.3619300  1.4648684

Variable: Percent of female borrowers (%)
  cluster_means[[i]]   t_value     p_value   Mean_est   CI_lower  CI_upper
1          0.6208452  3.269287 0.006097636  0.6208452  0.2105862 1.0311042
2         -0.4914893 -1.538700 0.147856287 -0.4914893 -1.1815509 0.1985723
3         -0.2012203 -1.300045 0.229788852 -0.2012203 -0.5581423 0.1557017


Variable: Profit margin (%)
  cluster_means[[i]]      t_value     p_value      Mean_est   CI_lower   CI_upper
1      -0.5833010858 -2.052796197 0.060789331 -0.5833010858 -1.1971688 0.03056666
2       0.5827259727  3.749760833 0.002428583  0.5827259727  0.2469971 0.91845483
3       0.0008946203  0.002791969 0.997840700  0.0008946203 -0.7380098 0.73979899

Variable: Yield on gross portfolio (real) (%)
  cluster_means[[i]]    t_value      p_value   Mean_est   CI_lower   CI_upper
1          0.2611467  0.7966291 4.399732e-01  0.2611467 -0.4470539  0.9693474
2         -0.6087448 -6.3899728 2.380663e-05 -0.6087448 -0.8145537 -0.4029359
3          0.5407081  1.8705692 9.832094e-02  0.5407081 -0.1258671  1.2072834

Variable: Number of loans outstanding, Credit Products , Enterprise Finance, Loans To Small And Medium Enterprises (%)
  cluster_means[[i]]   t_value      p_value   Mean_est   CI_lower   CI_upper
1         -0.2908690 -5.371109 0.0001272932 -0.2908690 -0.4078624 -0.1738756
2          0.4360062  1.055492 0.3104219921  0.4360062 -0.4564061  1.3284185
3         -0.2257690 -5.897212 0.0003628841 -0.2257690 -0.3140520 -0.1374859

Variable: Number of loans outstanding, Methodology, Individual (%)
  cluster_means[[i]]   t_value      p_value   Mean_est   CI_lower   CI_upper
1         -1.0148033 -8.381505 1.337884e-06 -1.0148033 -1.2763732 -0.7532335
2          0.4895958  2.251917 4.225762e-02  0.4895958  0.0199037  0.9592878
3          0.8169896  7.118261 1.001719e-04  0.8169896  0.5523208  1.0816583
```

D2: Coefficients, Standard Errors & Significance Multinomial Logistic Regression model (via R-studio)



```
> print(results)
  Coefficients..Intercept. Coefficients..Gross.loan.portfolio.to.total.assets.....
2                -8.183042                                                8.518188
3                -5.048803                                                5.591531
  Coefficients..Loan.loss.rate..... Coefficients..Percent.of.female.borrowers.....
2                        -18.200108                                       -16.3940
3                          6.303989                                       -16.7913
  Coefficients..Profit.margin..... Coefficients..Yield.on.gross.portfolio..real......
2                        12.820691                                         -36.015191
3                         6.732389                                           7.765363
  Coefficients..Number.of.loans.outstanding..Credit.Products...Enterprise.Finance..Loans.To.Small.And.Medium.Enterprises.....
2                                                                                                                   3.4349324
3                                                                                                                  -0.8072735
  Coefficients..Number.of.loans.outstanding..Methodology..Individual.....
2                                                                20.93046
3                                                                21.95658
  Standard.Errors..Intercept. Standard.Errors..Gross.loan.portfolio.to.total.assets.....
2                    9675.046                                                   2252.221
3                   70586.255                                                  33862.088
  Standard.Errors..Loan.loss.rate..... Standard.Errors..Percent.of.female.borrowers.....
2                             35546.63                                          3962.434
3                             23346.29                                         18797.279
  Standard.Errors..Profit.margin..... Standard.Errors..Yield.on.gross.portfolio..real......
2                            1033.519                                              4897.394
3                           36595.224                                              7912.012
  Standard.Errors..Number.of.loans.outstanding..Credit.Products...Enterprise.Finance..Loans.To.Small.And.Medium.Enterprises.....
2                                                                                                                       35524.26
3                                                                                                                      230755.30
  Standard.Errors..Number.of.loans.outstanding..Methodology..Individual.....
2                                                                   5410.562
3                                                                   9689.170
  Z.values..Intercept. Z.values..Gross.loan.portfolio.to.total.assets.....
2        -8.457884e-04                                        0.0037821285
3        -7.152672e-05                                        0.0001651266

  Z.values..Loan.loss.rate..... Z.values..Percent.of.female.borrowers.....
2                 -0.0005120065                              -0.0041373566
3                  0.0002700211                              -0.0008932838
  Z.values..Profit.margin..... Z.values..Yield.on.gross.portfolio..real......
2                 0.0124048878                                  -0.0073539495
3                 0.0001839691                                   0.0009814651
  Z.values..Number.of.loans.outstanding..Credit.Products...Enterprise.Finance..Loans.To.Small.And.Medium.Enterprises.....
2                                                                                                            9.669258e-05
3                                                                                                           -3.498397e-06
  Z.values..Number.of.loans.outstanding..Methodology..Individual..... P.values..Intercept.
2                                                         0.003868444            0.9993252
3                                                         0.002266095            0.9999429
  P.values..Gross.loan.portfolio.to.total.assets..... P.values..Loan.loss.rate.....
2                                           0.9969823                     0.9995915
3                                           0.9998682                     0.9997846
  P.values..Percent.of.female.borrowers..... P.values..Profit.margin.....
2                                  0.9966989                    0.9901026
3                                  0.9992873                    0.9998532
  P.values..Yield.on.gross.portfolio..real......
2                                      0.9941325
3                                      0.9992169
  P.values..Number.of.loans.outstanding..Credit.Products...Enterprise.Finance..Loans.To.Small.And.Medium.Enterprises.....
2                                                                                                               0.9999229
3                                                                                                               0.9999972
  P.values..Number.of.loans.outstanding..Methodology..Individual..... Significance..Intercept.
2                                                           0.9969134
3                                                           0.9981919
  Significance..Gross.loan.portfolio.to.total.assets..... Significance..Loan.loss.rate.....
2
3
  Significance..Percent.of.female.borrowers..... Significance..Profit.margin.....
2
3
  Significance..Yield.on.gross.portfolio..real......
2
3
```



```
  Significance..Number.of.loans.outstanding..Credit.Products...Enterprise.Finance..Loans.To.Small.And.Medium.Enterprises.....
2
3
  Significance..Number.of.loans.outstanding..Methodology..Individual.....
2
3
```

Appendix E : ROBUSTNESS CHECK RESULTS

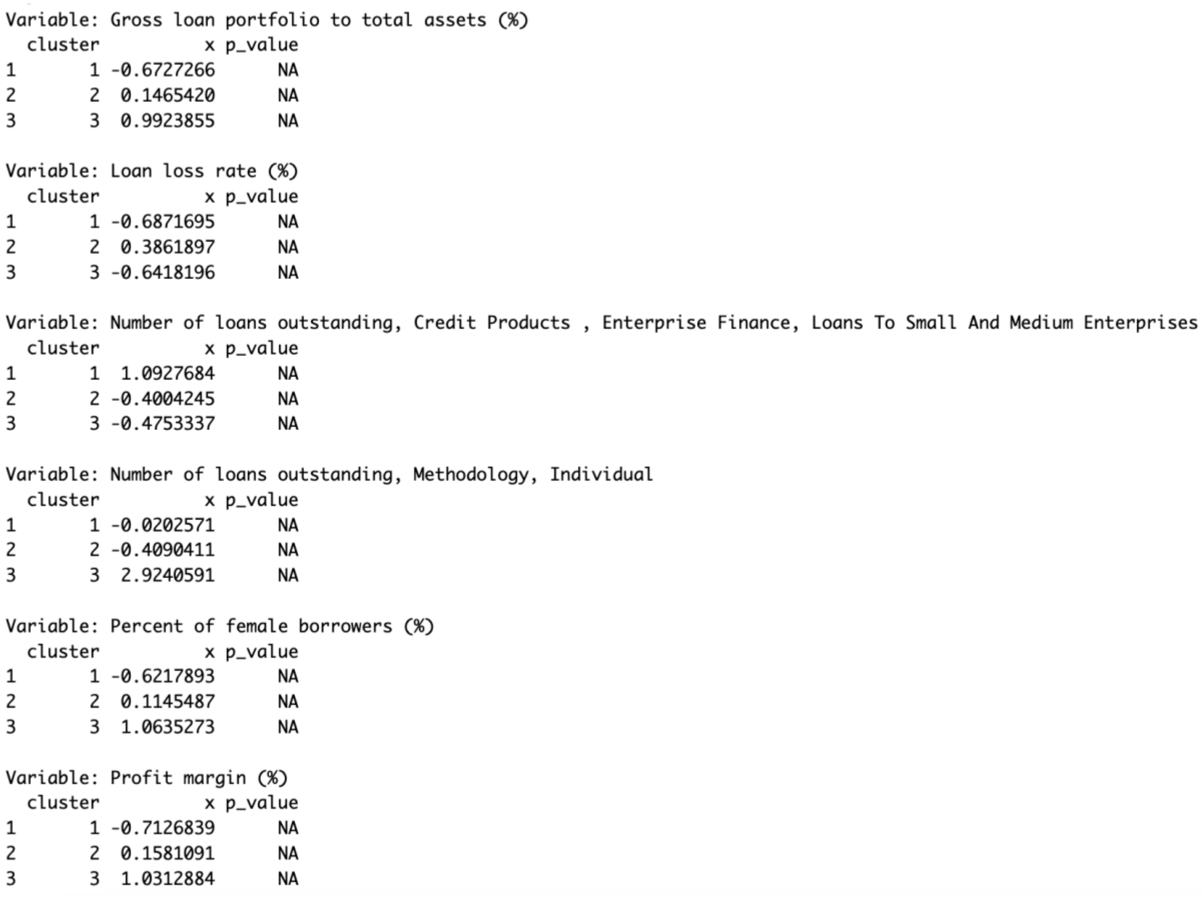
E1: Results table

TABLE 3 ROBUSTNESS CLUSTER RESULTS, MULTINOMIAL LOGISTIC REGRESSION

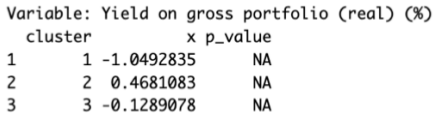
	1:	**2:**		**3:**	
	Cluster mean (normalized)	Cluster mean (normalized)	Coefficient (b's)	Cluster mean (normalized)	Coefficient (b's)
Intercept			-4.2601 (0.9999)		39.7137 (0.9986)
Loan Loss Rate (*Lossrate*)	-0.6872 (NA)*	0.3862	3.7307 (0.9999)	-0.6418	-64.0861 (0.9999)
Gross loan portfolio to total assets (%) (*ratio_GrLoan/TA)*	-0.6727	0.1465	8.7675 (0.9999)	0.9924	28.6146 (0.9999)
Percent of female borrowers (%) (*%femaleborrowers*)	-0.6218	0.1145	-23.3562 (0.9998)	1.0635	-28.9501 (0.9923)
Profit margin (%) (*profitmargin*)	-0.7127	0.1581	63.0169 (0.9991)	-1.0493	-14.0090 (0.9972)
Yield on gross portfolio (real) (%) (*realyield*)	-1.0493	0.4681	25.3714 (0.9998)	-0.1289	1.7999 (0.9999)
Number of loans outstanding, Credit Products, Enterprise Finance, loans to SME (%) (*%SME*)	1.0928	-0.4004	-5.2712 (0.9999)	-0.4753	-35.1604 (0.9990)
Number of loans outstanding, Methodology, Individual (%) (*%Individual*)	-0.0203	-0.4090	-48.4221 (0.9996)	2.9241	51.8554 (0.9962)
#Observations	1	6		4	

Note: The column labelled cluster mean shows the estimated mean of a variable within the cluster. Between the parentheses is the t-value of the difference between the cluster mean the full sample mean (i.e. over all clusters). The coefficients b's show how the log odds of being in cluster 2 or 3 instead of being in the reference cluster 1 changes when the variable increases by one unit. The parentheses *p <0,10; **p<0,05; ***p<0,01 indicate the significance levels of the regression standard errors of respectively 10, 5 and 1 per cent. The variable of most interest is *Lossrate;* the other variables are *ratio_GrLoan/TA*, *%femaleborrowers, profitmargin, realyield, %SME* and *%Individual. (NA)* All cluster means have an NA of p-value, since there are too little observations to give a p-value.*

E2: Means of clusters

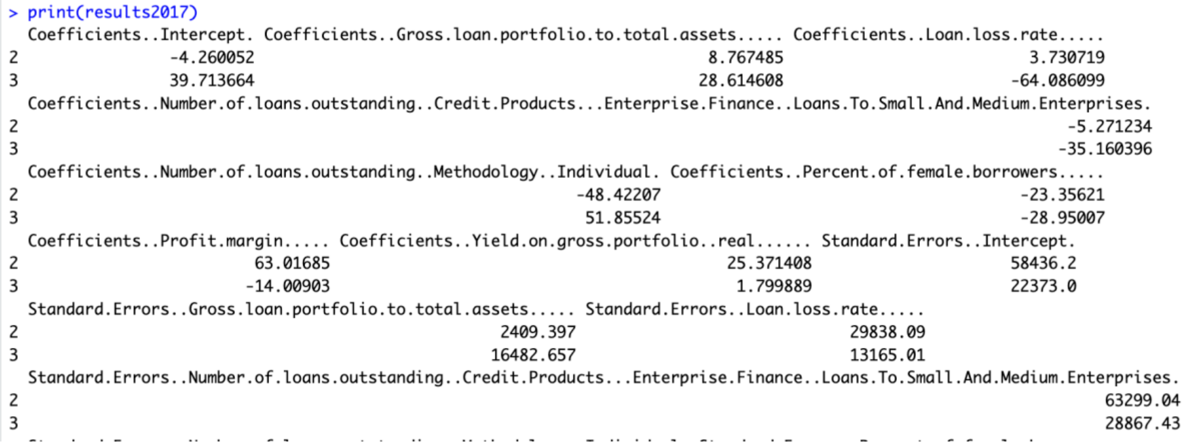


```
Variable: Gross loan portfolio to total assets (%)
  cluster          x p_value
1       1 -0.6727266      NA
2       2  0.1465420      NA
3       3  0.9923855      NA

Variable: Loan loss rate (%)
  cluster          x p_value
1       1 -0.6871695      NA
2       2  0.3861897      NA
3       3 -0.6418196      NA

Variable: Number of loans outstanding, Credit Products , Enterprise Finance, Loans To Small And Medium Enterprises
  cluster          x p_value
1       1  1.0927684      NA
2       2 -0.4004245      NA
3       3 -0.4753337      NA

Variable: Number of loans outstanding, Methodology, Individual
  cluster          x p_value
1       1 -0.0202571      NA
2       2 -0.4090411      NA
3       3  2.9240591      NA

Variable: Percent of female borrowers (%)
  cluster          x p_value
1       1 -0.6217893      NA
2       2  0.1145487      NA
3       3  1.0635273      NA

Variable: Profit margin (%)
  cluster          x p_value
1       1 -0.7126839      NA
2       2  0.1581091      NA
3       3  1.0312884      NA
```

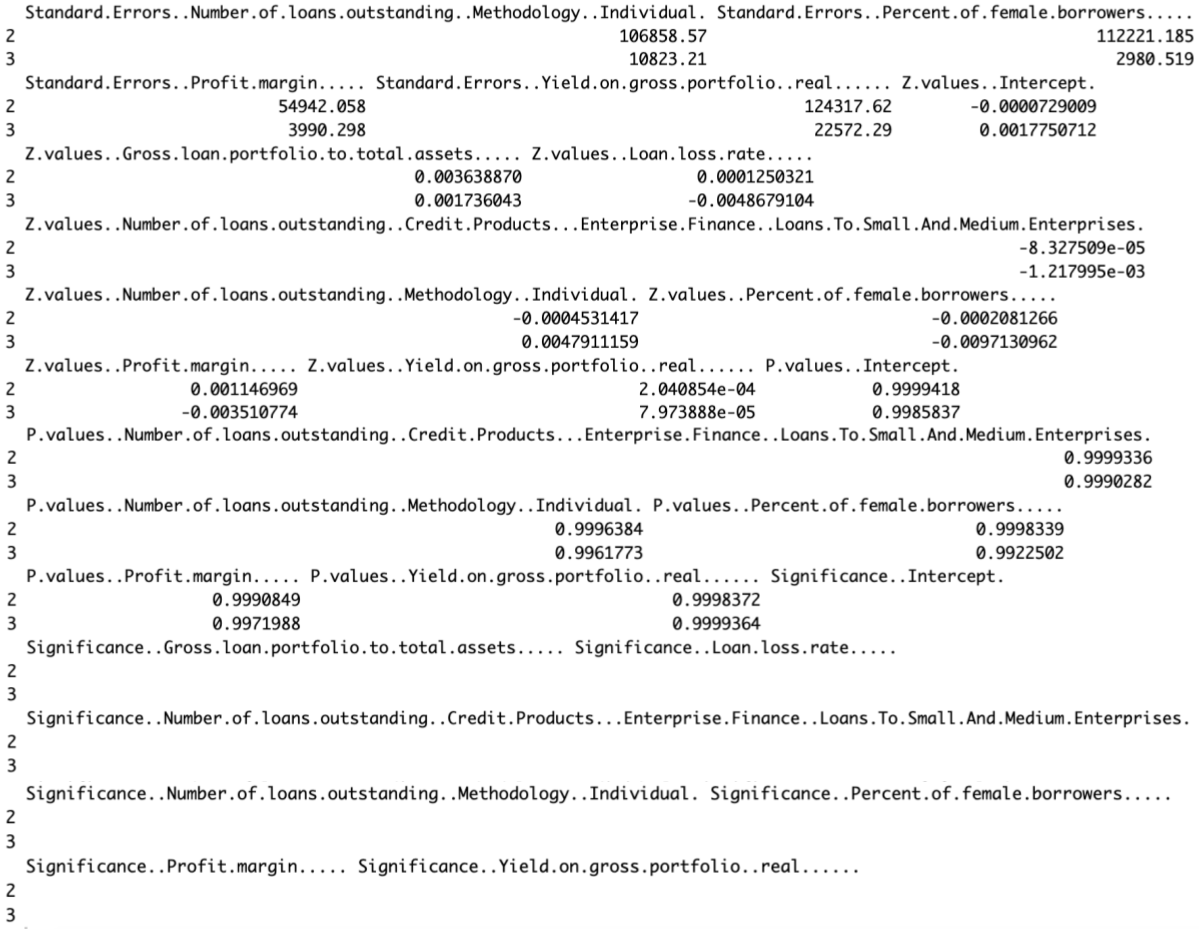


```
Variable: Yield on gross portfolio (real) (%)
  cluster          x p_value
1       1 -1.0492835      NA
2       2  0.4681083      NA
3       3 -0.1289078      NA
```

E3: Multinomial Logistic Regression Analysis



```
> print(results2017)
  Coefficients..Intercept. Coefficients..Gross.loan.portfolio.to.total.assets..... Coefficients..Loan.loss.rate.....
2                -4.260052                                                8.767485                          3.730719
3                39.713664                                               28.614608                        -64.086099
  Coefficients..Number.of.loans.outstanding..Credit.Products...Enterprise.Finance..Loans.To.Small.And.Medium.Enterprises.
2                                                                                                                -5.271234
3                                                                                                               -35.160396
  Coefficients..Number.of.loans.outstanding..Methodology..Individual. Coefficients..Percent.of.female.borrowers.....
2                                                           -48.42207                                      -23.35621
3                                                            51.85524                                      -28.95007
  Coefficients..Profit.margin..... Coefficients..Yield.on.gross.portfolio..real...... Standard.Errors..Intercept.
2                         63.01685                                          25.371408                     58436.2
3                        -14.00903                                           1.799889                     22373.0
  Standard.Errors..Gross.loan.portfolio.to.total.assets..... Standard.Errors..Loan.loss.rate.....
2                                                   2409.397                             29838.09
3                                                  16482.657                             13165.01
  Standard.Errors..Number.of.loans.outstanding..Credit.Products...Enterprise.Finance..Loans.To.Small.And.Medium.Enterprises.
2                                                                                                                   63299.04
3                                                                                                                   28867.43
```



```
  Standard.Errors..Number.of.loans.outstanding..Methodology..Individual. Standard.Errors..Percent.of.female.borrowers.....
2                                                              106858.57                                        112221.185
3                                                               10823.21                                          2980.519
  Standard.Errors..Profit.margin..... Standard.Errors..Yield.on.gross.portfolio..real...... Z.values..Intercept.
2                           54942.058                                             124317.62        -0.0000729009
3                            3990.298                                              22572.29         0.0017750712
  Z.values..Gross.loan.portfolio.to.total.assets..... Z.values..Loan.loss.rate.....
2                                         0.003638870                  0.0001250321
3                                         0.001736043                 -0.0048679104
  Z.values..Number.of.loans.outstanding..Credit.Products...Enterprise.Finance..Loans.To.Small.And.Medium.Enterprises.
2                                                                                                       -8.327509e-05
3                                                                                                       -1.217995e-03
  Z.values..Number.of.loans.outstanding..Methodology..Individual. Z.values..Percent.of.female.borrowers.....
2                                                   -0.0004531417                              -0.0002081266
3                                                    0.0047911159                              -0.0097130962
  Z.values..Profit.margin..... Z.values..Yield.on.gross.portfolio..real...... P.values..Intercept.
2                  0.001146969                                   2.040854e-04            0.9999418
3                 -0.003510774                                   7.973888e-05            0.9985837
  P.values..Number.of.loans.outstanding..Credit.Products...Enterprise.Finance..Loans.To.Small.And.Medium.Enterprises.
2                                                                                                           0.9999336
3                                                                                                           0.9990282
  P.values..Number.of.loans.outstanding..Methodology..Individual. P.values..Percent.of.female.borrowers.....
2                                                       0.9996384                                  0.9998339
3                                                       0.9961773                                  0.9922502
  P.values..Profit.margin..... P.values..Yield.on.gross.portfolio..real...... Significance..Intercept.
2                    0.9990849                                      0.9998372
3                    0.9971988                                      0.9999364
  Significance..Gross.loan.portfolio.to.total.assets..... Significance..Loan.loss.rate.....
2
3
  Significance..Number.of.loans.outstanding..Credit.Products...Enterprise.Finance..Loans.To.Small.And.Medium.Enterprises.
2
3
  Significance..Number.of.loans.outstanding..Methodology..Individual. Significance..Percent.of.female.borrowers.....
2
3
  Significance..Profit.margin..... Significance..Yield.on.gross.portfolio..real......
2
3
```

Appendix F : R-STUDIO DO-FILE

```
##Final R Studio File Master Thesis UU Viënna van Holsteijn 4666240

##Open MIX1
## install.packages("reshape2")
library (reshape2)
library(tidyr)
library(dplyr)
library(cluster)

##Create Dataset

year_to_keep <- "YR2017 [YR2017]"

Poging111 <- MIX1 %>%
  pivot_longer(cols = c("YR2019 [YR2019]", "YR2018 [YR2018]", "YR2017 [YR2017]", "YR2016 [YR2016]", "YR2015 [YR2015]"),
               names_to = "Year",
               values_to = "Value") %>%
  filter(Year == year_to_keep) %>%
  group_by(`Country Code`, `Series Name`) %>%
  summarize(Value = first(Value), .groups = "drop") %>%
  pivot_wider(names_from = "Series Name", values_from = "Value")

### CLEAN DATA

year_to_keep <- "YR2018 [YR2018]"

Poging2018 <- MIX1 %>%
  pivot_longer(cols = c("YR2019 [YR2019]", "YR2018 [YR2018]", "YR2017 [YR2017]", "YR2016 [YR2016]", "YR2015 [YR2015]"),
               names_to = "Year",
               values_to = "Value") %>%
  filter(Year == year_to_keep) %>%
  group_by(`Country Code`, `Series Name`, `Level`,`Country Name`) %>%
  summarize(Value = first(Value), .groups = "drop") %>%
  pivot_wider(names_from = "Series Name", values_from = "Value")

Poging2018[Poging2018 == '..'] <- NA

missing_count <- rowSums(is.na(Poging2018))
data_filtered <- Poging2018 %>%
  filter(missing_count < 55)

##66 observations, 85 variables

##DELETE UNNECESSARY COLUMNS

columns_to_delete <- c("Clients below poverty line (%)", "Number of loans outstanding, Credit Products , Household Financing, Other
household finance", "Number of loans outstanding, Credit Products , Household Financing, Mortgage/housing", "Number of loans
outstanding, Credit Products , Household Financing, Consumption", "Number of loans outstanding, Relationship, External Customers",
"Number of loans outstanding, Methodology, Village Banking SHG", "Number of loans outstanding, Relationship, Management And Staff",
"Write offs","Number of loans outstanding, Credit Products , Enterprise Finance, Large Corporations" , "Borrower retention rate (%)",
"Personnel expense / loan portfolio (%)", "Yield on gross portfolio (nominal) (%)", "Number of new borrowers" , "Percent of female
managers (%)" ,"Percent of female staff (%)" ,"Average salary / GNI per capita", "Number of loans outstanding, Gender, Legal Entity",
"Number of loans outstanding, Gender, Male", "Number of loans outstanding, Gender, Female", "NA", "Last Updated: 01/30/2023" ,"Data
from database: MIX Market", "Value of transactions, Delivery channels, Mobile banking", "Value of transactions, Delivery channels,
Internet", "Value of transactions, Delivery channels, ATMs", "Personnel expense / assets (%)", "Percentage of total transactions by mobile
banking, value (%)", "Percentage of total transactions by mobile banking, number (%)", "Percentage of total transactions by internet,
number (%)", "Percentage of total transactions by internet, value (%)","Percentage of total transactions at ATMs, value (%)","Percentage
of total transactions at roving staff, value (%)", "Percentage of total transactions at roving staff, number (%)", "Percentage of total
transactions at sub-branches, value (%)", "Percentage of total transactions at sub-branches, number (%)","Average outstanding balance",
"Clients below poverty line", "Cost per loan", "Gross Loan Portfolio", "Interest income on loan portfolio","Number of loans
outstanding","Donations", "Education services outreach", "Enterprise services outreach", "Net loan portfolio","Number of active
borrowers","Number of loans disbursed","Number of enterprises financed","Offices","Operational self sufficiency (%)","Percent of female
board members (%)","Percentage of total transactions at ADCs, number (%)","Percentage of total transactions at ADCs, value
(%)","Percentage of total transactions at agents, number (%)","Percentage of total transactions at agents, value (%)","Percentage of total
transactions at ATMs, number (%)","Percentage of total transactions at ATMs, value","Percentage of total transactions at merchant POS,
number (%)","Percentage of total transactions at merchant POS, value (%)")
```

```
# Delete columns with specific names
V2018_filtered <- Poging2018[, !names(Poging2018) %in% columns_to_delete]

V2018_filtered[V2018_filtered == '..'] <- NA

missing_count <- rowSums(is.na(V2018_filtered))
data_filtered <- V2018_filtered %>%
  filter(missing_count < 7)
##57 observations, 28 variables

# Count the number of NA values in each column
na_counts <- sapply(data_filtered, function(x) sum(is.na(x)))

# Print the results
print(na_counts)

###Delete columns with little observations:
###Delete columns with less than 47 observations:
# Step 1: Count the number of non-missing observations in each column
non_missing_counts <- colSums(!is.na(data_filtered))

# Step 2: Subset the dataset to keep only columns with 47 or more non-missing observations
data_filtered <- data_filtered[, non_missing_counts >= 47]

# View the updated dataset
str(data_filtered)
###57 observations, 26 variables (Women Empowerment services outreach & Methodolody, Solidarity Group deleted)

###MAKE EVERYTHING NUMERIC
# Check the structure of the data to identify columns to convert
str(data_filtered)
cols_to_convert <- c('Average loan balance per borrower','Yield on gross portfolio (real) (%)', 'Total Equity', 'Staff turnover rate (%)', 'Risk
coverage (%)' ,'Retained earnings', 'Profit margin (%)' ,'Percent of female borrowers (%)', 'Average number of active borrowers', 'Average
outstanding balance / GNI per capita (%)', 'Cost per borrower','Debt to equity ratio','Gross loan portfolio to total assets (%)','Loan loss rate
(%)', 'Net Income before taxes and donations', 'Average number of loans outstanding', 'Number of loans outstanding, Credit Products ,
Enterprise Finance, Loans To Small And Medium Enterprises', 'Number of loans outstanding, Credit Products , Enterprise Finance'
,'Number of loans outstanding, Credit Products , Enterprise Finance, Microenterprise' , 'Number of loans outstanding, Credit Products ,
Household Financing', 'Number of loans outstanding, Location, Rural', 'Number of loans outstanding, Location, Urban', 'Number of loans
outstanding, Methodology, Individual')

# Convert specified columns to numeric
data_filtered[cols_to_convert] <- data_filtered[cols_to_convert] %>%
  lapply(function(x) as.numeric(as.character(x)))

# Check the structure again to confirm the conversion
str(data_filtered)

# Generate a summary of the numeric columns
summary(data_filtered[, cols_to_convert])

####
###Add columns from type of loans to percentages
# Identify the columns to be created by matching patterns
columns_to_create <-
names(data_filtered)[grepl("Number.of.loans.outstanding|Number.of.new.borrowers|Credit.Products|Enterprise.Finance|Household.Fin
ancing|Location|Methodology|Relationship", names(data_filtered))]

# Ensure the 'Average number of loans outstanding' column is numeric
data_filtered <- data_filtered %>%
  mutate(`Average number of loans outstanding` = as.numeric(`Average number of loans outstanding`))

# Loop through columns to create new columns
for (col in columns_to_create) {
  if(col != "Average number of loans outstanding") {  # Ensure not to divide the column by itself

    # Convert the column to numeric
    data_filtered <- data_filtered %>%
      mutate(!!sym(col) := as.numeric(!!sym(col)))

    # Create the new column name
    new_col_name <- paste0(gsub("\\.", " ", col), " (%)")
    # Perform the division and create the new column
    data_filtered <- data_filtered %>%
      mutate(!!new_col_name := !!sym(col) / `Average number of loans outstanding`)
  }
}

# View the resulting dataframe
print(data_filtered)

##REMOVE COLUMNS WITH ACTUAL NUMBERS
# List of columns to remove
columns_to_remove <- c(
  "Number of loans outstanding, Credit Products , Enterprise Finance",
  "Number of loans outstanding, Credit Products , Enterprise Finance, Loans To Small And Medium Enterprises",
  "Number of loans outstanding, Credit Products , Enterprise Finance, Microenterprise",
  "Number of loans outstanding, Credit Products , Household Financing",
  "Number of loans outstanding, Location, Rural",
  "Number of loans outstanding, Location, Urban",
  "Number of loans outstanding, Methodology, Individual"
)

# Remove the specified columns from the dataset
data_filtered <- data_filtered[, !colnames(data_filtered) %in% columns_to_remove]
##from 33 to 26 variables with 57 observations

###Cluster analysis

###STEP 1: PREPROCESS DATA
# Remove non-numeric variables
numeric_data <- data_filtered[, sapply(data_filtered, is.numeric)]

# Handle missing values
numeric_data <- na.omit(numeric_data)

# Scale the data
scaled_data <- scale(numeric_data)
###31 observations with 23 variables (3 other variables are not numeric, so removed, are informational variables)

##CORRELATION MATRIX
# Get the column names
column_names <- colnames(numeric_data)

# Print the column names
print(column_names)

# Create the named vector for renaming with trimmed column names
new_names <- c(
  "Average loan balance per borrower" = "AVGloan",
  "Average number of active borrowers" = "AVGborrowers",
  "Average number of loans outstanding" = "AVGnumberloans",
  "Average outstanding balance / GNI per capita (%)" = "ratio_balance/GNI",
  "Cost per borrower" = "costborrower",
  "Debt to equity ratio" = "ratio_D/E",
  "Gross loan portfolio to total assets (%)" = "ratio_GrLoan/TA",
  "Loan loss rate (%)" = "Lossrate",
  "Net Income before taxes and donations" = "NIBT&donations",
  "Percent of female borrowers (%)" = "%femaleborrowers",
  "Profit margin (%)" = "profitmargin",
  "Retained earnings" = "retainedearnings",
  "Risk coverage (%)" = "riskcov",
  "Staff turnover rate (%)" = "staffturnover",
  "Total Equity" = "TotalEq",
  "Yield on gross portfolio (real) (%)" = "realyield",
  "Number of loans outstanding, Credit Products , Enterprise Finance (%)" = "%Enterprisefinance",
  "Number of loans outstanding, Credit Products , Enterprise Finance, Loans To Small And Medium Enterprises (%)" = "%SME",
  "Number of loans outstanding, Credit Products , Enterprise Finance, Microenterprise (%)" = "%Microenterprise",
  "Number of loans outstanding, Credit Products , Household Financing (%)" = "%Householdfinance",
  "Number of loans outstanding, Location, Rural (%)" = "%Rural",
  "Number of loans outstanding, Location, Urban (%)" = "%Urban",
  "Number of loans outstanding, Methodology, Individual (%)" = "%Individual",
```

```r
  "Number of loans outstanding, Methodology, Individual" = "%Individual",
  "Number of loans outstanding, Credit Products , Enterprise Finance, Loans To Small And Medium Enterprises" = "%SME"
)

# Trim any leading or trailing whitespace in column names
colnames(numeric_data) <- trimws(colnames(numeric_data))

# Rename the columns using a loop to ensure exact matching
for (old_name in names(new_names)) {
  if (old_name %in% colnames(numeric_data)) {
    colnames(numeric_data)[colnames(numeric_data) == old_name] <- new_names[old_name]
  }
}

# Verify the renaming
print(colnames(numeric_data))

# Assuming numeric_data is your dataset with numeric variables
cor_matrix <- cor(numeric_data)

# Print the correlation matrix with new column names
print(cor_matrix)

# Install the corrplot package
#install.packages("corrplot")

# Load the corrplot package
library(corrplot)

# Visualize the correlation matrix
corrplot(cor_matrix, method = "pie", type = "lower", tl.cex = 0.8, tl.srt = 45)
#corrplot(cor_matrix, method = "number", type = "lower", tl.cex = 0.8, tl.srt = 45)

# Define the correlation threshold
threshold <- 0.6

# Compute the correlation matrix
cor_matrix <- cor(numeric_data)

# Find the pairs of variables with absolute correlation higher than the threshold
high_cor_pairs <- which(abs(cor_matrix) > threshold, arr.ind = TRUE)

# Exclude self-correlations
high_cor_pairs <- high_cor_pairs[high_cor_pairs[, 1] != high_cor_pairs[, 2], ]

# Identify variables to remove (keep the first variable, remove the second in each pair)
variables_to_remove <- unique(colnames(numeric_data)[high_cor_pairs[, 2]])

# Ensure 'realyield' is not removed
variables_to_remove <- setdiff(variables_to_remove, "realyield")

# Print the variables to remove
print(variables_to_remove)

# Remove the highly correlated variables
numeric_data_filtered <- numeric_data[, !colnames(numeric_data) %in% variables_to_remove]

# Add 'realyield' back into the filtered dataset if it was initially removed
if (!"realyield" %in% colnames(numeric_data_filtered)) {
  numeric_data_filtered <- cbind(numeric_data_filtered, numeric_data["realyield"])
}

# Verify the new dataset
print(colnames(numeric_data_filtered))

# Visualize the updated correlation matrix
cor_matrix_filtered <- cor(numeric_data_filtered)
corrplot(cor_matrix_filtered, method = "number", tl.cex = 0.9, tl.srt = 45)

#######
####
```

```
###VARIABLES AFTER CORRELATION
print(colnames(numeric_data_filtered))

##DELETE UNNECESSARY COLUMNS
columns_to_delete <- c( "Women empowerent services outreach" ,"Number of loans outstanding, Methodology, Solidarity Group"
,"Number of loans outstanding, Location, Urban" ,"Number of loans outstanding, Location, Rural" ,"Number of loans outstanding, Credit
Products , Household Financing" ,'Number of loans outstanding, Credit Products , Enterprise Finance, Microenterprise', 'Number of loans
outstanding, Credit Products , Enterprise Finance' , "Total Equity" ,"Risk coverage (%)" ,"Retained earnings" , "Net Income before taxes and
donations", "Cost per borrower" ,"Average outstanding balance / GNI per capita (%)" , "Average number of active borrowers", "Average
loan balance per borrower", "Clients below poverty line (%)", "Number of loans outstanding, Credit Products , Household Financing, Other
household finance", "Number of loans outstanding, Credit Products , Household Financing, Mortgage/housing", "Number of loans
outstanding, Credit Products , Household Financing, Consumption", "Number of loans outstanding, Relationship, External Customers",
"Number of loans outstanding, Methodology, Village Banking SHG", "Number of loans outstanding, Relationship, Management And Staff",
"Write offs","Number of loans outstanding, Credit Products , Enterprise Finance, Large Corporations" , "Borrower retention rate (%)",
"Personnel expense / loan portfolio (%)", "Yield on gross portfolio (nominal) (%)", "Number of new borrowers" , "Percent of female
managers (%)" ,"Percent of female staff (%)" ,"Average salary / GNI per capita", "Number of loans outstanding, Gender, Legal Entity",
"Number of loans outstanding, Gender, Male", "Number of loans outstanding, Gender, Female", "NA", "Last Updated: 01/30/2023" ,"Data
from database: MIX Market", "Value of transactions, Delivery channels, Mobile banking", "Value of transactions, Delivery channels,
Internet", "Value of transactions, Delivery channels, ATMs", "Personnel expense / assets (%)", "Percentage of total transactions by mobile
banking, value (%)", "Percentage of total transactions by mobile banking, number (%)", "Percentage of total transactions by internet,
number (%)", "Percentage of total transactions by internet, value (%)","Percentage of total transactions at ATMs, value (%)","Percentage
of total transactions at roving staff, value (%)", "Percentage of total transactions at roving staff, number (%)", "Percentage of total
transactions at sub-branches, value (%)", "Percentage of total transactions at sub-branches, number (%)","Average outstanding balance",
"Clients below poverty line", "Cost per loan", "Gross Loan Portfolio", "Interest income on loan portfolio","Number of loans
outstanding","Donations", "Education services outreach", "Enterprise services outreach", "Net loan portfolio","Number of active
borrowers","Number of loans disbursed","Number of enterprises financed","Offices","Operational self sufficiency (%)","Percent of female
board members (%)","Percentage of total transactions at ADCs, number (%)","Percentage of total transactions at ADCs, value
(%)","Percentage of total transactions at agents, number (%)","Percentage of total transactions at agents, value (%)","Percentage of total
transactions at ATMs, number (%)","Percentage of total transactions at ATMs, value","Percentage of total transactions at merchant POS,
number (%)","Percentage of total transactions at merchant POS, value (%)")

# Delete columns with specific names
V2018_filtered <- Poging2018[, !names(Poging2018) %in% columns_to_delete]

V2018_filtered[V2018_filtered == '..'] <- NA

missing_count <- rowSums(is.na(V2018_filtered))
data_filtered <- V2018_filtered %>%
  filter(missing_count < 1)
###37 observations with 12 variables (should be 13 with average number of loans outstanding)

#CREATE % in SME & INDIVIDUAL
####
###Add columns from type of loans to percentages
# Identify the columns to be created by matching patterns
columns_to_create <-
names(data_filtered)[grepl("Number.of.loans.outstanding|Number.of.new.borrowers|Credit.Products|Enterprise.Finance|Household.Fin
ancing|Location|Methodology|Relationship", names(data_filtered))]

# Ensure the 'Average number of loans outstanding' column is numeric
data_filtered <- data_filtered %>%
  mutate(`Average number of loans outstanding` = as.numeric(`Average number of loans outstanding`))

# Loop through columns to create new columns
for (col in columns_to_create) {
  if(col != "Average number of loans outstanding") {  # Ensure not to divide the column by itself

    # Convert the column to numeric
    data_filtered <- data_filtered %>%
      mutate(!!sym(col) := as.numeric(!!sym(col)))

    # Create the new column name
    new_col_name <- paste0(gsub("\\.", " ", col), " (%)")

    # Perform the division and create the new column
    data_filtered <- data_filtered %>%
      mutate(!!new_col_name := !!sym(col) / `Average number of loans outstanding`)
  }
}

# View the resulting dataframe
```

```
print(data_filtered)

##REMOVE COLUMNS WITH ACTUAL NUMBERS
# List of columns to remove
columns_to_remove <- c(
  "Number of loans outstanding, Credit Products , Enterprise Finance, Loans To Small And Medium Enterprises",
  "Number of loans outstanding, Methodology, Individual",
  "Average number of loans outstanding"
)

# Remove the specified columns from the dataset
data_filtered <- data_filtered[, !colnames(data_filtered) %in% columns_to_remove]
##37 observations with 12 variables

###CLUSTER ANALYSIS

###MAKE EVERYTHING NUMERIC
# Check the structure of the data to identify columns to convert
str(data_filtered)
cols_to_convert <- c('Yield on gross portfolio (real) (%)', 'Staff turnover rate (%)' , 'Profit margin (%)' ,'Percent of female borrowers (%)',
'Debt to equity ratio','Gross loan portfolio to total assets (%)','Loan loss rate (%)', 'Number of loans outstanding, Credit Products ,
Enterprise Finance, Loans To Small And Medium Enterprises (%)', 'Number of loans outstanding, Methodology, Individual (%)')

# Convert specified columns to numeric
data_filtered[cols_to_convert] <- data_filtered[cols_to_convert] %>%
  lapply(function(x) as.numeric(as.character(x)))

# Check the structure again to confirm the conversion
str(data_filtered)

# Generate a summary of the numeric columns
summary(data_filtered[, cols_to_convert])

###STEP 1: PREPROCESS DATA
# Remove non-numeric variables
numeric_data <- data_filtered[, sapply(data_filtered, is.numeric)]

# Handle missing values (if any)
numeric_data <- na.omit(numeric_data)

# Scale the data (optional but recommended)
scaled_data <- scale(numeric_data)

###STEP 2: CHOOSING NUMBER OF CLUSTERS
###ELBOW
# Set a random seed for reproducibility
set.seed(123)

# Determine the optimal number of clusters using the elbow method
wss <- numeric(10)

for (i in 1:10) {
  # Run k-means clustering with multiple starts
  kmeans_result <- kmeans(scaled_data, centers = i, nstart = 25)

  # Sum of within-cluster sum of squares
  wss[i] <- sum(kmeans_result$withinss)
}

# Plot the within sum of squares for each number of clusters
plot(1:10, wss, type = "b", xlab = "Number of Clusters", ylab = "Within Sum of Squares")

###NO CLEAR ELBOW, SO DO SILHOUETTE SCORES

##SILHOUETTE SCORES
# Set a random seed for reproducibility
set.seed(123)

# Calculate silhouette scores for k values from 2 to 10
silhouette_scores <- sapply(2:10, function(k) {
  # Run k-means clustering
  km <- kmeans(scaled_data, centers = k, nstart = 25)

  # Compute the silhouette scores
  ss <- silhouette(km$cluster, dist(scaled_data))

  # Return the average silhouette width
  mean(ss[, "sil_width"])
})

# Plot the silhouette scores
plot(2:10, silhouette_scores, type = "b", xlab = "Number of Clusters", ylab = "Average Silhouette Width")

###NUMBER OF CLUSTERS = 3 = highest ASW

###STEP 3: Perform Cluster Analysis
# Perform k-means clustering with the optimal number of clusters
num_clusters <- 3

cluster_model <- kmeans(scaled_data, centers = num_clusters)

###STEP 4: Analyze Clusters
# Get cluster assignments
cluster_assignments <- cluster_model$cluster
scaled_data <- as.data.frame(scaled_data)
print(cluster_assignments)

scaled_data$cluster_assignments <- c(1, 1, 1, 2, 2, 1, 2, 1, 1, 1, 1, 2, 1, 3, 2, 3, 1, 1, 3, 3, 1, 2, 2, 1, 2, 3, 2, 3, 2, 2, 2, 3, 3, 2, 3, 2, 1)

str(scaled_data)

###NEWWW
#CLUSTER 1 =  14 observations
#CLUSTER 2 =  14 observations
#CLUSTER 3 =  9 observations
#TOTAL 37 observations

####Check clusters

###NEW
# Assuming scaled_data is your dataset with 37 observations and cluster_assignments

# Calculate cluster means
cluster_means <- aggregate(. ~ cluster_assignments, data = scaled_data, FUN = mean)

# Assuming cluster_means is your dataset with cluster means and the correct column names
# Verify the structure of cluster_means
print(cluster_means)

# Define the column names
column_names <- c(
  "Debt to equity ratio",
  "Gross loan portfolio to total assets (%)",
  "Loan loss rate (%)",
  "Percent of female borrowers (%)",
  "Profit margin (%)",
  "Staff turnover rate (%)",
  "Yield on gross portfolio (real) (%)",
  "Number of loans outstanding, Credit Products , Enterprise Finance, Loans To Small And Medium Enterprises (%)",
  "Number of loans outstanding, Methodology, Individual (%)"
)

# Set the column names of cluster_means (excluding the first column which is cluster_assignments)
colnames(cluster_means)[-1] <- column_names

# Reshape data from wide to long format manually
cluster_means_long <- data.frame(
  cluster_assignments = rep(cluster_means$cluster_assignments, each = length(column_names)),
  Variable = rep(column_names, times = nrow(cluster_means)),
  Mean = as.vector(as.matrix(cluster_means[, -1]))
)
```

```r
# Print the reshaped data to verify
print(cluster_means_long)

# Adjust margins (bottom, left, top, right) to fit the rotated x-axis labels
par(mar = c(12, 4, 4, 2))

# Create the plot
# Plotting setup
plot.new()
plot.window(xlim = c(1, length(column_names)), ylim = range(cluster_means_long$Mean))
axis(1, at = 1:length(column_names), labels = column_names, las = 2, cex.axis = 0.5)  # Smaller labels
axis(2)
box()
title(main = "Cluster Means for Each Variable", xlab = "", ylab = "Mean Value")  # Removed y-axis title

# Define colors for clusters
colors <- c("red", "blue", "green")

# Add points to the plot
for (i in 1:nrow(cluster_means)) {
  points(1:length(column_names), cluster_means[i, -1], col = colors[cluster_means$cluster_assignments[i]], pch = 16)
}

# Add legend with adjusted position to the right of the plot
legend("topright", inset = c(-0.05, 0), legend = paste(unique(cluster_means$cluster_assignments)),
       col = colors, pch = 16, xpd = TRUE)

##DECISION: LEAVE VARIABLE STAFF TURNOVER RATE OUT! IS TOO MUCH IN COMMON WITH LOAN LOSS RATE
#DECISION: LEAVE D/E RATIO OUT! TOO MUCH IN COMMON WITH YIELD GROSS PORTFOLIO (REAL)!

###REMOVE VARIABLES
# Remove the two specified variables from scaled_data
scaled_data <- scaled_data[, !(colnames(scaled_data) %in% c("Staff turnover rate (%)", "Debt to equity ratio"))]

# Verify the removal
print(colnames(scaled_data))

#install.packages("nnet")
library(nnet)

column_names <- c(
  'Gross loan portfolio to total assets (%)',
  'Loan loss rate (%)',
  'Percent of female borrowers (%)',
  'Profit margin (%)',
  'Yield on gross portfolio (real) (%)',
  'Number of loans outstanding, Credit Products , Enterprise Finance, Loans To Small And Medium Enterprises (%)',
  'Number of loans outstanding, Methodology, Individual (%)',
  'cluster_assignments'
)

###MULTINOMIAL LOGISTIC REGRESSION ANALYSIS

set.seed(123)  # For reproducibility
colnames(scaled_data) <- column_names

# Fit the multinomial logistic regression model
model <- multinom(cluster_assignments ~ ., data = scaled_data)

# Extract coefficients and standard errors
coefs <- summary(model)$coefficients
std_err <- summary(model)$standard.errors

# Calculate z-values and p-values
z_values <- coefs / std_err
p_values <- 2 * (1 - pnorm(abs(z_values)))

# Create a function to add significance stars
significance_stars <- function(p_values) {
  stars <- rep("", length(p_values))
  stars[p_values < 0.1] <- "."
  stars[p_values < 0.05] <- "*"
  stars[p_values < 0.01] <- "**"
  stars[p_values < 0.001] <- "***"
  return(stars)
}

# Add stars to p-values
stars <- apply(p_values, 2, significance_stars)

# Combine the results into a data frame for easy viewing
results <- data.frame(
  Coefficients = coefs,
  `Standard Errors` = std_err,
  `Z-values` = z_values,
  `P-values` = p_values,
  `Significance` = stars
)

print(results)


###CLUSTER MEANS WITH SIGNIFICANCE

# Define your data and variables of interest
your_data <- scaled_data

variables_of_interest <- c(
  'Gross loan portfolio to total assets (%)',
  'Loan loss rate (%)',
  'Percent of female borrowers (%)',
  'Profit margin (%)',
  'Yield on gross portfolio (real) (%)',
  'Number of loans outstanding, Credit Products , Enterprise Finance, Loans To Small And Medium Enterprises (%)',
  'Number of loans outstanding, Methodology, Individual (%)'
)

cluster_column <- "cluster_assignments"

# Compute overall sample mean for each variable
overall_means <- sapply(variables_of_interest, function(variable) mean(your_data[[variable]]))

# Compute cluster means for each variable
cluster_means <- lapply(variables_of_interest, function(variable) {
  tapply(your_data[[variable]], your_data[[cluster_column]], mean)
})

# Perform two-sided t-tests for each variable and add results to cluster_means
for (i in seq_along(variables_of_interest)) {
  variable <- variables_of_interest[i]
  # Initialize an empty data frame to store t-test results
  test_results <- data.frame(Cluster = integer(), Mean = numeric(), t_value = numeric(), p_value = numeric(), Mean_est = numeric(),
CI_lower = numeric(), CI_upper = numeric(), stringsAsFactors = FALSE)
  for (cluster in names(cluster_means[[i]])) {
    # Perform t-test for each cluster vs overall mean
    t_test <- t.test(your_data[[variable]][your_data[[cluster_column]] == as.numeric(cluster)], mu = overall_means[i])
    # Extract and format results
    test_result <- data.frame(Cluster = as.integer(cluster),
                              Mean = cluster_means[[i]][[cluster]],
                              t_value = t_test$statistic,
                              p_value = t_test$p.value,
                              Mean_est = t_test$estimate,
                              CI_lower = t_test$conf.int[1],
                              CI_upper = t_test$conf.int[2])
    # Append to test_results data frame
    test_results <- rbind(test_results, test_result)
  }
  # Assign results to cluster_means list
  cluster_means[[i]] <- cbind(cluster_means[[i]], test_results[, c("t_value", "p_value", "Mean_est", "CI_lower", "CI_upper")])
}

# Print or use cluster_means as needed
```

```r
for (i in seq_along(variables_of_interest)) {
  variable <- variables_of_interest[i]
  cat("Variable:", variable, "\n")
  print(cluster_means[[i]])
  cat("\n")
}

#########
###ROBUSTNESS MODEL CHECK
###YEAR 2017
### YEAR 2017 DATA PREPARATION ###

# Define the year to keep
year_to_keep <- "YR2017 [YR2017]"

# Filter and pivot the data similar to 2018
ROBUST2017 <- MIX1 %>%
  pivot_longer(cols = c("YR2019 [YR2019]", "YR2018 [YR2018]", "YR2017 [YR2017]", "YR2016 [YR2016]", "YR2015 [YR2015]"),
               names_to = "Year",
               values_to = "Value") %>%
  filter(Year == year_to_keep) %>%
  group_by(`Country Code`, `Series Name`, `Level`, `Country Name`) %>%
  summarize(Value = first(Value), .groups = "drop") %>%
  pivot_wider(names_from = "Series Name", values_from = "Value")

# Replace '..' with NA
ROBUST2017[ROBUST2017 == '..'] <- NA

# Filter out rows with too many missing values
missing_count <- rowSums(is.na(ROBUST2017))
ROBUST2017 <- ROBUST2017 %>%
  filter(missing_count < 55)  # Adjust as per your criteria

# View the structure of ROBUST2017
str(ROBUST2017)

### DATA CLEANING AND PREPARATION ###
# Assuming you need to remove unnecessary columns as done previously in 2018

# List of columns to delete
columns_to_delete <- c(
  "Clients below poverty line (%)",
  "Number of loans outstanding, Credit Products , Household Financing, Other household finance",
  "Number of loans outstanding, Credit Products , Household Financing, Mortgage/housing",
  "Number of loans outstanding, Credit Products , Household Financing, Consumption",
  "Number of loans outstanding, Relationship, External Customers",
  "Number of loans outstanding, Methodology, Village Banking SHG",
  "Number of loans outstanding, Relationship, Management And Staff",
  "Write offs",
  "Number of loans outstanding, Credit Products , Enterprise Finance, Large Corporations",
  "Borrower retention rate (%)",
  "Personnel expense / loan portfolio (%)",
  "Yield on gross portfolio (nominal) (%)",
  "Number of new borrowers",
  "Percent of female managers (%)",
  "Percent of female staff (%)",
  "Average salary / GNI per capita",
  "Number of loans outstanding, Gender, Legal Entity",
  "Number of loans outstanding, Gender, Male",
  "Number of loans outstanding, Gender, Female",
  "NA",
  "Last Updated: 01/30/2023",
  "Data from database: MIX Market",
  "Value of transactions, Delivery channels, Mobile banking",
  "Value of transactions, Delivery channels, Internet",
  "Value of transactions, Delivery channels, ATMs",
  "Personnel expense / assets (%)",
  "Percentage of total transactions by mobile banking, value (%)",
  "Percentage of total transactions by mobile banking, number (%)",
  "Percentage of total transactions by internet, number (%)",
  "Percentage of total transactions by internet, value (%)",
  "Percentage of total transactions at ATMs, value (%)",
  "Percentage of total transactions at roving staff, value (%)",
  "Percentage of total transactions at roving staff, number (%)",
  "Percentage of total transactions at sub-branches, value (%)",
  "Percentage of total transactions at sub-branches, number (%)",
  "Average outstanding balance",
  "Clients below poverty line",
  "Cost per loan",
  "Gross Loan Portfolio",
  "Interest income on loan portfolio",
  "Number of loans outstanding",
  "Donations",
  "Education services outreach",
  "Enterprise services outreach",
  "Net loan portfolio",
  "Number of active borrowers",
  "Number of loans disbursed",
  "Number of enterprises financed",
  "Offices",
  "Staff turnover rate (%)",
  "Debt to equity ratio",
  "Operational self sufficiency (%)",
  "Percent of female board members (%)",
  "Percentage of total transactions at ADCs, number (%)",
  "Percentage of total transactions at ADCs, value (%)",
  "Percentage of total transactions at agents, number (%)",
  "Percentage of total transactions at agents, value (%)",
  "Percentage of total transactions at ATMs, number (%)",
  "Percentage of total transactions at ATMs, value",
  "Percentage of total transactions at merchant POS, number (%)",
  "Percentage of total transactions at merchant POS, value (%)"
)


#install.packages("nnet")
library(nnet)

# Delete unnecessary columns from ROBUST2017
ROBUST2017 <- ROBUST2017[, !names(ROBUST2017) %in% columns_to_delete]

# Remove rows with missing values (adjust as needed)
ROBUST2017 <- ROBUST2017[complete.cases(ROBUST2017), ]

# View the structure of cleaned ROBUST2017
str(ROBUST2017)

### MAKE ALL VARIABLES NUMERIC ###
# Check the structure of the data to identify columns to convert
str(ROBUST2017)

# Define columns to convert to numeric
cols_to_convert <- c(
  'Yield on gross portfolio (real) (%)',
  'Profit margin (%)',
  'Percent of female borrowers (%)',
  'Gross loan portfolio to total assets (%)',
  'Loan loss rate (%)',
  'Number of loans outstanding, Credit Products , Enterprise Finance, Loans To Small And Medium Enterprises',
  'Number of loans outstanding, Methodology, Individual'
)

# Convert specified columns to numeric
ROBUST2017[cols_to_convert] <- ROBUST2017[cols_to_convert] %>%
  lapply(function(x) as.numeric(as.character(x)))

# Check the structure again to confirm the conversion
str(ROBUST2017)

# Generate a summary of the numeric columns
summary(ROBUST2017[, cols_to_convert])
```

```
### STEP 1: PREPROCESS DATA ###
# Remove non-numeric variables
numeric_data_2017 <- ROBUST2017[, sapply(ROBUST2017, is.numeric)]

# Handle missing values (if any)
numeric_data_2017 <- na.omit(numeric_data_2017)

# Scale the data (optional but recommended)
scaled_data_2017 <- scale(numeric_data_2017)

### STEP 3: Perform Cluster Analysis ###
# Perform k-means clustering with the optimal number of clusters
num_clusters <- 3

cluster_model_2017 <- kmeans(scaled_data_2017, centers = num_clusters)

### STEP 4: Analyze Clusters ###
# Get cluster assignments
cluster_assignments_2017 <- cluster_model_2017$cluster
scaled_data_2017 <- as.data.frame(scaled_data_2017)
print(cluster_assignments_2017)

# Load necessary library for multinomial logistic regression
library(nnet)

# View the column names of ROBUST2017
column_names <- colnames(scaled_data_2017)

# Print the column names
print(column_names)

set.seed(123)  # For reproducibility
ROBUST2017 <- data.frame(
  matrix(runif(11 * length(column_names)), nrow = 11, ncol = length(column_names))
)
colnames(ROBUST2017) <- column_names

# Fit the multinomial logistic regression model
model <- multinom(cluster_assignments_2017 ~ ., data = ROBUST2017)

# Extract coefficients and standard errors
coefs <- summary(model)$coefficients
std_err <- summary(model)$standard.errors

# Calculate z-values and p-values
z_values <- coefs / std_err
p_values <- 2 * (1 - pnorm(abs(z_values)))

# Create a function to add significance stars
significance_stars <- function(p_values) {
  stars <- rep("", length(p_values))
  stars[p_values < 0.1] <- "."
  stars[p_values < 0.05] <- "*"
  stars[p_values < 0.01] <- "**"
  stars[p_values < 0.001] <- "***"
  return(stars)
}

# Add stars to p-values
stars <- apply(p_values, 2, significance_stars)

# Combine the results into a data frame for easy viewing
results2017 <- data.frame(
  Coefficients = coefs,
  `Standard Errors` = std_err,
  `Z-values` = z_values,
  `P-values` = p_values,
  `Significance` = stars
)

print(results2017)
```

```
###MEANS + SIGNIFICANCE 2017
###CLUSTER MEANS WITH SIGNIFICANCE

# Example assuming 'scaled_data_2017' is your dataframe and has been cleaned

# Define variables of interest
variables_of_interest <- c(
  'Gross loan portfolio to total assets (%)',
  'Loan loss rate (%)',
  'Number of loans outstanding, Credit Products , Enterprise Finance, Loans To Small And Medium Enterprises',
  'Number of loans outstanding, Methodology, Individual',
  'Percent of female borrowers (%)',
  'Profit margin (%)',
  'Yield on gross portfolio (real) (%)'
)

# Calculate cluster means for each variable
cluster_means <- lapply(variables_of_interest, function(variable) {
  aggregate(scaled_data_2017[[variable]], by = list(cluster = scaled_data_2017$cluster_assignments), FUN = mean)
})

# Perform statistical testing (compute p-values)
for (i in seq_along(variables_of_interest)) {
  variable <- variables_of_interest[i]
  overall_mean <- mean(scaled_data_2017[[variable]])

  # Perform t-test for each cluster vs overall mean if there are enough observations
  cluster_means[[i]]$p_value <- sapply(cluster_means[[i]]$x, function(x) {
    cluster_data <- scaled_data_2017[[variable]][scaled_data_2017$cluster_assignments == x]

    # Check if there are enough observations for the t-test
    if (length(cluster_data) >= 2) {
      t_test <- t.test(cluster_data, mu = overall_mean)$p.value
    } else {
      t_test <- NA  # Set to NA if not enough observations
    }
    return(t_test)
  })
}

# View results (print or inspect cluster means and p-values)
for (i in seq_along(variables_of_interest)) {
  variable <- variables_of_interest[i]
  cat("Variable:", variable, "\n")
  print(cluster_means[[i]])
  cat("\n")
}
###P-values are NA due to low number of observations

# Define your data and variables of interest
your_data <- scaled_data_2017

variables_of_interest <- c(
  'Gross loan portfolio to total assets (%)',
  'Loan loss rate (%)',
  'Number of loans outstanding, Credit Products , Enterprise Finance, Loans To Small And Medium Enterprises',
  'Number of loans outstanding, Methodology, Individual',
  'Percent of female borrowers (%)',
  'Profit margin (%)',
  'Yield on gross portfolio (real) (%)'
)

cluster_column <- "cluster_assignments"

# Compute overall sample mean for each variable
overall_means <- sapply(variables_of_interest, function(variable) mean(your_data[[variable]]))

# Compute cluster means for each variable
cluster_means_2017 <- lapply(variables_of_interest, function(variable) {
  tapply(your_data[[variable]], your_data[[cluster_column]], mean)
})

# Perform two-sided t-tests for each variable and add results to cluster_means
for (i in seq_along(variables_of_interest)) {
 variable <- variables_of_interest[i]
 # Initialize an empty data frame to store t-test results
 test_results <- data.frame(Cluster = integer(), Mean = numeric(), t_value = numeric(), p_value = numeric(), Mean_est = numeric(),
CI_lower = numeric(), CI_upper = numeric(), stringsAsFactors = FALSE)
 for (cluster in names(cluster_means[[i]])) {
  # Perform t-test for each cluster vs overall mean
  t_test <- t.test(your_data[[variable]][your_data[[cluster_column]] == as.numeric(cluster)], mu = overall_means[i])
  # Extract and format results
  test_result <- data.frame(Cluster = as.integer(cluster),
                            Mean = cluster_means[[i]][[cluster]],
                            t_value = t_test$statistic,
                            p_value = t_test$p.value,
                            Mean_est = t_test$estimate,
                            CI_lower = t_test$conf.int[1],
                            CI_upper = t_test$conf.int[2])
  # Append to test_results data frame
  test_results <- rbind(test_results, test_result)
 }
 # Assign results to cluster_means list
 cluster_means[[i]] <- cbind(cluster_means[[i]], test_results[, c("t_value", "p_value", "Mean_est", "CI_lower", "CI_upper")])
}

# Print or use cluster_means as needed
for (i in seq_along(variables_of_interest)) {
 variable <- variables_of_interest[i]
 cat("Variable:", variable, "\n")
 print(cluster_means[[i]])
 cat("\n")
}


###END TEST ROBUSTNESS

############################

####Summary Descriptives
##General summary
summary(numeric_data)

##Add Standard Deviation
# Custom summary function to include mean, standard deviation, min, and max
custom_summary <- function(x) {
 c(
  Mean = mean(x, na.rm = TRUE),
  SD = sd(x, na.rm = TRUE),
  Min = min(x, na.rm = TRUE),
  Max = max(x, na.rm = TRUE)
 )
}

# Apply the custom summary function to all numeric columns in numeric_data
numeric_columns <- sapply(numeric_data, is.numeric)
summary_statistics <- sapply(numeric_data[, numeric_columns], custom_summary)

# Transpose the result for better readability
summary_statistics <- t(summary_statistics)

# Print the summary statistics
print(summary_statistics)

####FIGURES DESCRIPTIVE DATA
###HISTOGRAMS
# Create a histogram of the Loan loss rate (%)
hist(numeric_data$`Lossrate`,
     main = "Histogram of Loan Loss Rate (%)",
     xlab = "Loan Loss Rate (%)",
     ylab = "Frequency",
     col = "lightblue",
     border = "black")


#Create histogram for Percent of female borrowers (%):
hist(numeric_data$`%femaleborrowers`,
     main = "Histogram of Percent of female borrowers (%)",
     xlab = "Percent of female borrowers (%)",
     ylab = "Frequency",
     col = "lightblue",
     border = "black")

#Create histogram for Yield on gross portfolio (real) (%):
hist(numeric_data$`realyield`,
     main = "Histogram of Yield on gross portfolio (real) (%)",
     xlab = "Yield on gross portfolio (real) (%)",
     ylab = "Frequency",
     col = "lightblue",
     border = "black")

####THE END
```