Can We Predict Poverty Risk? A Data-Driven Approach to Identifying Vulnerable Populations and Providing Targeted Assistance

Abstract:

This project uses machine learning to predict poverty risk based on socioeconomic, demographic, and digital inclusion indicators. By analyzing data from 12,600 individuals across seven anonymized countries, we identify key vulnerability patterns across age groups. Our XGBoost model achieves high predictive accuracy and powers an interactive Shiny app that enables policymakers to simulate poverty risk and explore targeted interventions. We show that digital literacy, education, and rural residence significantly affect poverty outcomes. Our findings offer actionable, data-driven insights to support smarter poverty alleviation strategies and personalized policy design for vulnerable populations.

Course: 2024-2025 3 Data Science Lab (ECB3DSL)

Team: Group 1, Team 6

Aleksandra Tatko

Valeria Fuenmayor van Praag

Ana Sofia Polo Bleher

Supervisor(s): Tina Dulam

Word count: 4932 words

Utrecht University

Statement of Originality

We, the aforementioned students, herewith declare to have written this document and that we are responsible for the content of it. We declare that the text and the work presented in this document is original and that no sources other than those mentioned in the text and its references have been used in creating it.
Utrecht University School of Economics is responsible solely for the supervision of completion of the work, not for the content.

Disclosure Statement

In this project, we have made use of the following Generative AI tools:

Tool	**Purpose**
ChatGTP	*Correction of code*
Gamma	*Presentation layout*
Grammarly	*Text correction (grammar, spelling, syntax)*

Division of Work

We, the aforementioned students, herewith declare that we have divided the work on this project and this project paper as stated in the following table:

Section and Content	**Name and Student Number**
1 Introduction	All students
2 Data and Methodology	All students
3 Descriptive Analysis	All students
4 Discussion	All students
5 Limitations	All students
6 Conclusion	All students

Signatures

Ana Sofia Polo Bleher
a.s.polobleher@students.uu.nl
1461087

Valeria Fuenmayor van Praag
v.s.fuenmayorvanpraag@students.uu.nl
9858989

Aleksandra Tatko
a.tatko@students.uu.nl
8022399

Table of Content

1 Introduction

Poverty remains one of the most pressing global challenges, affecting millions and hindering development across various dimensions of society. As socioeconomic disparities widen, understanding and predicting poverty risk becomes increasingly essential. This study focuses on the intersection of poverty prediction, vulnerability assessment, and risk analysis, with the specific research question: *"Can we predict poverty risk using socioeconomic, demographic, and digital indicators, and how do these drivers differ across age groups? Additionally, how can this prediction be leveraged to create an interactive tool for policymakers and individuals to assess and mitigate poverty risk?"*

To develop effective strategies, the factors correlated with poverty must also be considered by the influence of policy changes (Sekhampu, 2013). A well-designed prediction tool can serve as an effective targeting mechanism for identifying households and individuals who are eligible for poverty reduction, development, and social programs (Cadena-Palacios et al., 2024). Our findings will also help to create a validation strategy to identify beneficiaries of poverty reduction, which could be developed further using platforms like Shiny.

The significance of this research lies in its potential to contribute to both societal welfare and academic discourse. By applying predictive modelling and machine learning to social issues, we can assess the feasibility of predicting vulnerability to poverty while uncovering such key drivers. Previous research indicates that the measurement of poverty is inherently difficult and controversial, as there is no universally accepted definition of the poverty line or threshold (Zeller, 2013). Instead of merely classifying individuals as poor or non-poor, we adopt a probabilistic approach that assesses the chance of being poor. This aligns with the broader concept of vulnerability, defined as the "probability of falling into poverty in the future" (Zhang & Wan, 2008).

A range of socioeconomic and demographic factors have been identified in prior research as significant determinants of poverty risk. Among these, age (Sekhampu, 2013; Iqbal & Awan, 2015; Maloma, 2016), employment status (Sekhampu, 2013; Maloma, 2016), education level of the household head (Geda et al, 2001; Maloma, 2016; Iqbal & Awan, 2015; Wu et al, 2024; Zixi, 2021; Pokhriyal & Jacques, 2017), and health status (measured as life expectancy at birth) (Pokhriyal, 2017; Wu et al, 2024) have all been found to negatively correlate with poverty. Conversely, engagement in agricultural activities (Geda et al., 2001), dependency within a

household (Iqbal & Awan, 2015), and household size (Sekhampu, 2013; Geda et al, 2001; Iqbal & Awan, 2015) are associated with a higher risk of poverty. Gender (Cadena-Palacios et al., 2024) also plays a crucial role, particularly among students, where women from rural areas with parents who have low educational attainment and ethnic minority backgrounds are disproportionately at risk.

Additionally, methodological considerations play a key role in poverty prediction. Research has highlighted the importance of selecting an appropriate vulnerability threshold, with 50% suggested as an optimal cutoff to improve predictive power (Zhang & Wan, 2008). Moreover, permanent income calculations have been found to be more reliable when incorporating weighted past income. The choice of poverty line also affects measurement accuracy, with higher poverty lines (e.g., $2 per day rather than $1) improving classification reliability. In terms of modelling, the choice between econometric techniques (such as OLS) and machine learning approaches (such as Lasso) is context-dependent (Verme, 2020), with Lasso often identified as the most accurate predictor depending on the chosen poverty threshold (Afzal & Newhouse, 2015).

This study aims to bridge the gap between academic theory and practical policy application, emphasizing human-cantered data science. Our research provides a unique exploration of how predictive analytics can improve social welfare strategies, thereby enhancing discussions around poverty alleviation and intervention techniques in existing literature. By examining the efficacy of predictive methodologies across different age groups and their policy implications, we seek to offer actionable insights for targeted interventions.

Our findings will highlight the efficacy of the proposed methodologies and their implications for policy and individual empowerment. We demonstrate that age, employment status, educational attainment, financial activity, and phone technology usage are significant predictors of poverty risk. Notably, irregular employment and lower education levels increase poverty risk, while financial engagement and digital access reduce it, highlighting the importance of financial literacy and technology in mitigating poverty.

To achieve these objectives, our methodology encompasses several key steps. In Chapter 2, we begin with data exploration and preparation. We then apply a machine learning model specifically designed for poverty risk prediction in Chapter 3, followed by the development of an interactive tool for policymakers in Chapter 5. This tool will allow users to input various

socioeconomic and demographic variables, assess poverty risk probabilities, identify high-risk groups, and gain geographic insights through scenario testing. Finally, we will conclude our findings with actionable policy recommendations in Chapter 6.

2 Data and Methodology

This section describes the data used in the study, the preprocessing and transformation steps applied to the raw data, and the modelling framework employed.

2.1 Data Description

For this study, the dataset Poverty Probability Index & Economic Indicators from Kaggle is utilized (Yiu, J.). The dataset includes the Poverty Probability Index (PPI), which was predicted for seven unknown countries using 60 variables. The countries are named A, C, D, F, G, I and J to maintain the anonymity of the original project. Individuals were asked questions related to socioeconomic and demographic factors in these countries, where the poverty line of the individuals has a $2.50/day threshold. The data consists of 12,600 observations collected from the Financial Inclusion Insights household surveys.

The dataset includes essential demographic attributes such as age, gender, and educational level, which provides valuable context for analyzing financial behaviour and economic conditions of an individual. For the education level variable, ranges from 0 to 3, the exact meaning of the values is unspecified. However, it can be inferred that 0 indicates an individual with no education, 1 represents individuals who have completed primary and middle school, 2 corresponds to individuals who have completed high school, and 3 represents individuals who have attended higher education.

Additionally, the dataset includes key numeracy skills, such as the ability to add, divide, and perform compound calculations, which highlights an individual's capacity to perform mathematical calculations that are essential for financial calculations. Information on the employment type, financial habits, and income sources of individuals offers further insight into the available job opportunities within the labour market, their financial situation and how they can support themselves and their families.

The dataset also provides information on phone usage, with values ranging from having no access to a phone or limited functionality (0) to continuous access to a highly advanced phone (3), representing different levels of phone accessibility. The primary variable, *poverty_probability*, quantifies the likelihood of an individual experiencing poverty, serving as a central measure in this analysis.

When downloading the dataset, three files were provided: **test_values.csv**, **train_values.csv, and train_labels.csv** datasets. The CSV files, represent the typical split in a supervised machine learning. For this study, the **train_values.csv** dataset was selected as the primary dataset for analysis and manipulation to create a complete dataset. This decision was made because the **train_values.csv** dataset can be combined with the **train_labels.csv** dataset, which contains the probability of an individual being in poverty which the testing dataset did not include. By combining the datasets, supervised learning can be conducted, allowing the model to predict patterns between socioeconomic and demographic indicators and poverty probability.

2.2 Data Preprocessing

The dataset used for this analysis was first imported from an Excel file into R. Upon loading, the initial preprocessing steps involved the removal of five columns due to the absence of data values. The dataset was then cleaned to ensure that any data that was missing from the rows and columns were removed for consistency in the data.

To prevent duplicate entries from affecting the analysis, the dataset was filtered to retain only distinct observations. Moreover, binary variables initially represented as logical values (true/false) were converted into numerical value, with true being represented as 1 and false as 0. This transformation was necessary to facilitate statistical modelling and machine learning techniques that require numerical input. A summary of the dataset was generated to examine key descriptive statistics, particularly focusing on the `age_group` variable. Based on this summary, the dataset was refined by categorizing individuals into three age groups; this allows us to see the vulnerability of different age groups and provide a targeted policy that is suitable. The age variable was divided into three categories: young (15–25 years), middle-aged (26–45 years), and older individuals (46 years and above). This grouping was based on quartile distributions, ensuring that each age category captured meaningful distinctions within the dataset.

Furthermore, the missing values in the target variable `poverty_probability` were filtered out. The *`set.seed()`* function was used to ensure reproducibility, and the *`createDataPartition()`* function was applied to divide the dataset. The dataset was named `poverty_1`, which was split into training (50%), validation (30%), and test sets (20%) to facilitate predictive modelling. The distribution of `poverty_probability` across these subsets was visualized using a histogram to confirm that the data was evenly distributed across the partitions.

2.3 Model Building and Evaluation

Our methodological framework incorporates both classical regression techniques and advanced machine learning models to comprehensively capture the relationships between the predictors and poverty probability (Verme, 2020).

2.3.1 Baseline Linear Model

Initially, we employed a baseline linear regression model. A systematic approach with the *`generate_formulas()`* function was used to generate numerous candidate models by varying combinations of predictors, resulting in 26,235 possible regression specifications.
The models were evaluated using the mean squared error (MSE) computed using the *`lm_mse()`* function on the validation set, and the best-performing linear model achieved an average MSE of approximately 0.06.

To evaluate the reliability of our baseline linear regression model, several diagnostic checks were performed. First, we performed 10-fold cross-validation using the caret package in R. This approach divides the dataset into 10 subsets, iteratively training the model on nine subsets while validating on the remaining one. The average performance metrics obtained were an RMSE of 0.246, an MAE of 0.203, and an R^2 of 0.289.

Although an R^2 of 0.289 indicates that the model explains roughly 29% of the variability in poverty probability, the low R^2 suggests that there may be additional important variables, interactions, or non-linear effects not captured by the current model. Nonetheless, the consistent prediction error across different folds demonstrates that the model is stable, and its performance is not overly dependent on a particular train–validation split.

Next, we evaluated multicollinearity in our baseline linear regression model using the Variance Inflation Factor (VIF). All values for our predictors are substantially below the common threshold of 5 or 10, indicating that multicollinearity is not a concern in our model.

To assess the validity of our linear regression model, we further conducted a series of residual diagnostic tests. First, we examined the standard diagnostic plots—including Residuals vs. Fitted, Normal Q-Q, Scale-Location, and Residuals vs. Leverage—which raised concerns regarding non-constant variance. This observation was formally confirmed by the Breusch-Pagan test (BP = 800.84, df = 9, $p < 2.2e-16$), indicating significant heteroscedasticity.

To address this issue and ensure reliable inference, we computed heteroscedasticity-consistent (robust) standard errors using the HC1 estimator. The robust t-tests showed that all predictors remain highly significant.

These findings suggest that although the error variance is non-constant, robust inference confirms that the relationships between the predictors and the poverty probability are statistically reliable.

Final validations of our linear model include bootstrapping for coefficient stability where we conducted an ordinary nonparametric bootstrap with 1,000 replications. In each replication, the model was refitted on a resampled version of the combined validation and test dataset, and the coefficient estimates were recorded. The bootstrap results indicate that the original estimates are highly stable, as evidenced by minimal bias and narrow standard errors across all predictors. For instance, the coefficient for `countryF` was estimated at -0.25821 with a bootstrap standard error of approximately 0.01096 and a 95% percentile confidence interval of (-0.2807, -0.2359). Similarly, the intercept (0.89016) and the coefficients for `literacy` (-0.10880) and `num_financial_activities_last_year` (-0.03449) exhibited low bias and small standard errors, confirming the robustness of these estimates. Overall, the narrow confidence intervals and consistent bootstrap statistics reinforce that the relationships observed in our model are not sensitive to the sample used, thereby supporting the reliability of our findings for subsequent predictive analyses.

Lastly, we used Cook's distance to identify observations that might disproportionately affect our linear regression estimates. A substantial number of data points exceeded the threshold, indicating potential influence.

Therefore, we conducted a sensitivity analysis by refitting the model after excluding these flagged observations. The results showed that while individual coefficient estimates shifted slightly, the overall significance and direction of the predictors remained consistent. This suggests that no single group of outliers is driving our conclusions about the relationships between the predictors and poverty probability.

2.3.2 Regularization with LASSO Regression

Following the robustness checks on our baseline linear model, we employed LASSO regression to improve our predictive performance. First, we prepared our data by converting the training set into a model matrix and then used cross-validation to identify an optimal lambda value. With this optimal value, we refitted the LASSO model.

Our baseline linear regression model, after a series of robustness checks, achieved an R^2 of approximately 0.29 (with improvements to about 0.36 after removing influential outliers) and a corresponding MSE of 0.06 on the validation data. In contrast, our tuned LASSO model via cross-validation and refitted using the optimal lambda explains 37.71% of the variance. Moreover, the LASSO model produced a lower MSE of 0.055 on the validation set.

This improvement in both explained deviance and prediction error indicates that regularization not only enhances model parsimony by shrinking irrelevant coefficients to zero but also improves predictive performance,

2.3.3 Advanced Predictive Modeling with Ensemble Methods

To capture potential non-linear relationships, we evaluated two alternative methods. A Random Forest model with 500 trees (mtry = 2) achieved an RMSE of 0.2407 and explained 30.53% of the variance. In contrast, an XGBoost model (learning rate 0.1, max. depth 6, 100 rounds) yielded an RMSE of 0.2225 on the validation set, outperforming both the Random Forest and the LASSO model. In our final XGBoost model, variable importance is measured by the Gain metric, which quantifies each feature's contribution to reducing model error. By sorting this metric in descending order and selecting the top 10 features, we identified the most influential predictors of poverty probability. However, this ranking is specific to XGBoost and its hyperparameters; other models like LASSO or Random Forest may emphasize different predictors. XGBoost's importance values reflect complex, non-linear interactions, so the selected features reduce error effectively, but do not represent independent causal effects.

To assess the generalizability of our predictive model, we evaluated the final XGBoost model on an independent test set comprising 2383 observations. The model achieved a root mean squared error (RMSE) of 0.2230, a mean absolute error (MAE) of 0.1774, and an R^2 of 0.4156.

After conducting a grid search with 5-fold cross-validation on our training data, we identified optimal hyperparameters for our XGBoost model, which collectively minimized the out-of-sample error. We then retrained the XGBoost model using these best-tuned settings and evaluated it on the validation set. The final model achieved an RMSE of 0.2213 and an R^2 of 0.426 on the test set.

Given these improved results, XGBoost was selected as the final model for our project. Its robust performance underpins its integration into our interactive Shiny tool, which will enable policymakers and stakeholders to input socioeconomic and demographic variables and receive real-time predictions of poverty risk.

3 Descriptive Analyses

Understanding poverty probability requires analysing a combination of demographic, socioeconomic, and technological factors. We focus on identifying key predictors and exploring how they interact across youth, working-age adults, and older individuals. We selected the most suitable variables based on the economic literature about poverty as well as including the most predicting variables from our XGBoost model. The dataset consists of both categorical and numerical variables, each contributing to an understanding of poverty risk. This analysis provides an overview of the dataset's distributions, key variable relationships, and insights into poverty probability, with visual representations included in the Appendix.

3.1 Demographic Variables and Their Relationship to Poverty

3.1.1 Age

The age distribution (Figure 1) is right-skewed, with younger individuals forming the largest portion of the dataset. When segmented into three age groups (youth, working-age adults, and older adults), the poverty probability distribution (Figure 2) reveals important differences.

Contrary to conventional expectations, younger individuals exhibit the lowest poverty probability, while older individuals face the highest. This may reflect the financial dependency of younger individuals on family structures, whereas older adults may struggle with income insecurity, limited employment, and health-related costs, consistent with findings on aging and poverty in developing economies (Cadena-Palacios et al., 2024). The stark contrast across age groups underscores the need to target interventions according to life stage, and it highlights the relevance of incorporating age as a central feature in any predictive policy tool.

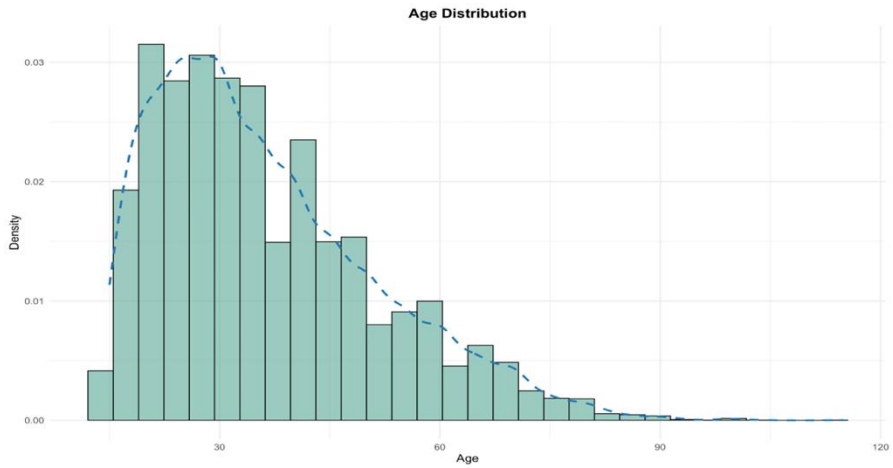


Figure 1. Age Distribution

Figure 2. Distribution of Poverty Probability by Age Group

3.1.2 Gender

The dataset consists of 55.8% females, making them only slightly the majority group (Figure 3). Poverty probability by gender (Appendix 1.2) shows that females have a slightly higher median poverty probability than males. This finding supports prior research indicating that

women face greater economic vulnerability due to wage gaps, caregiving responsibilities, and reduced employment opportunities (Cadena-Palacios et al., 2024). However, the distributions are similar, suggesting that gender alone is not a primary determinant of poverty risk but interacts with other socioeconomic factors.

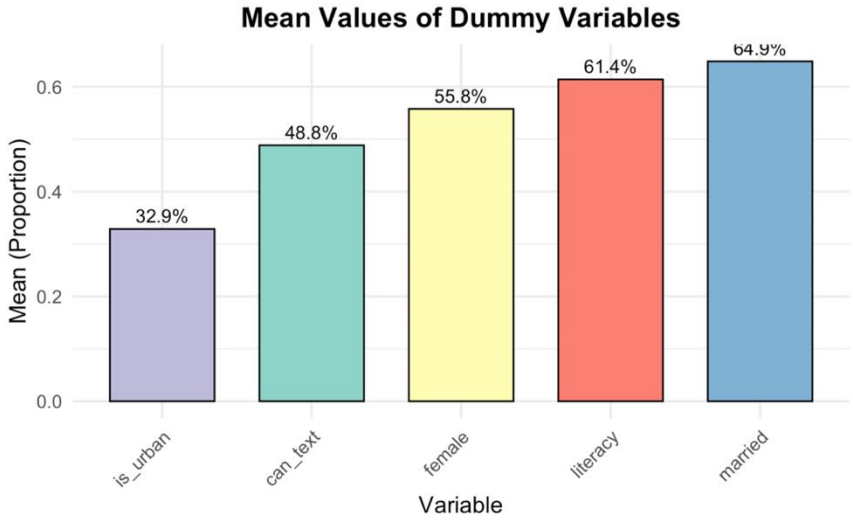


Figure 3. Mean Values of Selected Dummy Variables

3.1.3 Marital Status

Marriage is common in this dataset, with 64.9% of individuals being married (Figure 3). The poverty probability by marital status (Appendix 1.3) indicates that married individuals have a slightly higher median poverty probability than unmarried individuals, contradicting the assumption that marriage provides financial stability. One explanation is the economic burden of dependents, which increases household expenses, particularly in lower-income groups (Iqbal & Awan, 2015).

3.1.4 Urban vs. Rural Residence

Urban residents make up 32.9% of the sample, while the majority live in rural areas (Figure 3). The poverty probability by urban residence (Appendix 1.4) confirms that rural residents face significantly higher poverty risks, aligning with prior research showing that limited access to financial services, stable employment, and infrastructure disproportionately affects rural populations (Geda et al., 2001). However, urban residency alone does not guarantee economic security, as high living costs in cities contribute to financial precarity.

3.2 Socioeconomic Variables and Poverty

3.2.1 Education and Literacy

Education levels in the dataset are skewed toward lower attainment, with 36.81% of individuals having the lowest education level and only 9.59% achieving the highest level (Appendix 1.5). Poverty probability by education level (Appendix 1.6) shows that lower education levels correlate with higher poverty risks, reinforcing the widely documented relationship between education, employability, and income potential (Pokhriyal & Jacques, 2017).

Literacy is another key factor, with 61.4% of individuals being literate (Figure 3). The poverty probability by literacy status (Appendix 1.7) confirms that illiterate individuals face higher economic vulnerability, likely due to barriers in accessing financial services, job opportunities, and information necessary for economic mobility.

3.2.2 Relationship to Household Head

The dataset shows that 41.5% of individuals are household heads, 33.03% are spouses, and 17.28% are children (sons/daughters) (Appendix 1.8). The poverty probability by relationship to the household head (Appendix 1.9) reveals that non-head family members, such as children and dependents, have higher poverty risks, likely because they rely financially on household heads. This highlights the importance of household structure in shaping economic outcomes.

3.2.3 Employment

Employment type patterns reveal a high proportion of individuals who are not working or engaged in informal/self-employment (Appendix 1.10). Poverty probability by employment type (Appendix 1.11) shows that informal and irregular workers face the highest poverty risks, with salaried employees faring the best. These differences are particularly stark among working-age individuals, reinforcing the role of stable employment and formal sector integration in poverty reduction.

The employment status by age group (Appendix 1.12) also reveals that youth are more likely to be unemployed or informally employed, while older individuals are more often inactive or

retired. Despite low labour force participation among youth, their lower poverty probability further supports the notion of financial dependency on family rather than economic self-sufficiency, raising concerns about their vulnerability during transition into adulthood.

3.3 Digital and Financial Inclusion as Predictors of Poverty

3.3.1 Financial Literacy and Financial Activities

Financial engagement varies across the sample, with some individuals actively participating in multiple financial activities while others have limited involvement (Appendix 1.13). The poverty probability by the number of financial activities (Appendix 1.14) reveals that higher engagement in financial activities, such as savings and investments, correlates with lower poverty probability. This supports research indicating that financial literacy plays a crucial role in improving economic stability (Pokhriyal & Jacques, 2017).

3.3.2 Phone Technology and Texting Ability

Digital inclusion was analysed through two variables: phone technology usage and texting ability. A significant portion of individuals have access to mobile technology (Appendix 1.15). Poverty probability by phone technology usage (Appendix 1.16) demonstrates that higher levels of phone technology usage correspond with lower poverty risks, suggesting that digital access facilitates economic opportunities.

Similarly, individuals with texting ability ("can_text") exhibit lower poverty probabilities (Appendix 1.17). This suggests that mobile technology enables job searching, financial transactions, and digital literacy, reinforcing its role in economic empowerment. Our model confirmed that texting ability and financial literacy were the strongest predictors of poverty probability.

3.4 Correlation Insights

The correlation matrix (Figure 4) provides further insights into relationships among variables. Financial literacy is positively correlated with financial activities, confirming that financially

literate individuals are more engaged in economic decision-making. Texting ability correlates with financial literacy and phone technology usage, indicating that digital access enhances financial participation.

	is_urban	literacy	num_financial_activities_last_year	can_text	female	age	married	phone_technology
is_urban	1	0.19	0.17	0.2	0.01	-0.06	-0.09	0.22
literacy	0.19	1	0.29	0.45	-0.11	-0.26	-0.11	0.33
num_financial_activities_last_year	0.17	0.29	1	0.38	-0.08	-0.05	-0.09	0.44
can_text	0.2	0.45	0.38	1	-0.13	-0.29	-0.17	0.42
female	0.01	-0.11	-0.08	-0.13	1	-0.06	0.01	-0.2
age	-0.06	-0.26	-0.05	-0.29	-0.06	1	0.23	-0.13
married	-0.09	-0.11	-0.09	-0.17	0.01	0.23	1	-0.06
phone_technology	0.22	0.33	0.44	0.42	-0.2	-0.13	-0.06	1

Figure 4. Correlation Matrix with the highest predictors given by XGBoost model and variables found significant in literature.

Age is negatively correlated with financial activity engagement, suggesting that younger individuals participate less in financial systems, possibly due to financial dependence. Urban residency is weakly correlated with financial inclusion, highlighting that location alone does not determine financial literacy but interacts with other socioeconomic conditions.

4 Discussion

4.1 Shiny App

To effectively present our findings, we have developed a Shiny app that highlights key aspects of our project, with a particular focus on the Poverty Risk Predictor tab—an interactive tool designed to support policy-making and targeted assistance. This tool is intended for government agencies and policymakers, providing a data-driven approach to identify and address poverty risk across different age groups. By selecting key demographic and socio-economic variables, users can visualize how these factors influence poverty risk, with results displayed on an interactive graph. To enhance usability, we have color-coded different age groups, making it easier to analyse risk distributions and implement targeted interventions. Additionally, the graph includes three dashed reference lines that categorize poverty risk levels; low (0 to 0.25), medium (0.25 to 0.60) and high risk (0.60 to 1.00). This tool serves as a valuable resource for developing strategies to reduce poverty by enabling informed decision-making based on real-world data.

4.2 Policy Recommendations

4.2.1 Young (15- 26): Scholarship Programmes

Access to higher education is a key determinant in reducing poverty risk, particularly among young individuals aged 15–25. Socioeconomic disparities, especially in rural areas, along with limited educational opportunities, contribute significantly to the persistence of poverty. A well-designed scholarship programme should include; financial aid, academic resources, mentorship programs, and career guidance (Cadena-Palacios, C. N., 2024). This holistic approach ensures that students not only receive funding but also the necessary support to successfully complete their education and transition into the labour market. Studies indicate that educational scholarships improve completion rates among low-income youth, reduce poverty risk and increasing their chances of securing stable job. Scholarship programs can help bridge the gap between individuals who live in rural areas, are female and those who lack access to phone technology.

By offering students this program it will give them the opportunity to purse a high level of education in where they will be equipped with necessary skills, access to phone technology and help eliminate financial burden. While education is the first critical step it must be also be integrated with job training programs to ensure that the individuals can effectively apply their knowledge in the workforce. Governments can introduce job training and career development programs to further enhance employability (Wimer, C., 2020). This initiative can help individuals focus on skill development, job placement and career readiness. This ensures a seamless transition from education to the labour market, preventing employment gaps that could negatively impact economic stability.

4.2.2 Middle Age (26- 45): Community Literacy & Empowerment Workshops for Married Women

Among middle-aged individuals, our analysis reveals a significantly higher poverty risk for illiterate, married women—a trend that aligns with findings from Cadena-Palacios et al. (2024) and Maloma (2016). In many low-income and rural contexts, married women carry disproportionate household responsibilities, often without access to education or stable income. Illiteracy compounds this vulnerability by limiting access to employment, social services, and financial literacy.

We propose implementing Community Literacy & Empowerment Workshops targeting married, illiterate women aged 26–45. These locally run workshops should focus on practical literacy (e.g., reading, writing, navigating digital tools), financial education, and basic rights awareness. Studies show that adult literacy programs significantly improve household welfare, labour force participation, and self-reliance (Pokhriyal & Jacques, 2017; Wu et al., 2024).

Delivered through community centres, mobile libraries, or NGO partnerships, these workshops would reduce informational asymmetry, boost confidence, and empower women to participate in local economies or advocate for family resources. When combined with digital inclusion (e.g., phone use), literacy training can break the intergenerational cycle of poverty by equipping women to support their children's education and manage household finances more effectively.

4.2.3 Old (46+): Digital Empowerment Hubs in Rural Areas

Older individuals in rural areas face the highest predicted poverty risks according to our model—driven by reduced employment participation, isolation, and digital exclusion. While younger groups benefit from familial support or employment integration, older adults often lack both. Our analysis shows a strong link between digital literacy (proxied by texting ability) and lower poverty probabilities, affirming findings from Pokhriyal & Jacques (2017) and Wu et al. (2024) on digital access reducing economic vulnerability.

We recommend implementing Digital Empowerment Hubs—community-based evening centres in rural areas offering training on basic phone, texting, and laptop use. These hubs could empower older residents to access social protection services, mobile banking, and digital health tools. Evidence shows that improving digital inclusion enhances financial decision-making and promotes well-being (Hohberg et al., 2018; Zeller, 2013). While not aimed at labour market re-entry, the goal is to reduce social exclusion and increase autonomy in managing resources.

To maximize impact, hubs should be free, locally staffed, and paired with awareness campaigns. This initiative aligns with global digital inclusion goals and supports a shift from reactive to preventive poverty interventions.

5 Limitations

This study presents several limitations that may affect the validity and interpretation of our findings on poverty prediction.

First, the dataset utilized for this analysis does not offer a description of its columns, which constrains our understanding of the underlying variables. As a result, there is potential for misinterpretation or assumptions regarding variable meaning, which could introduce biases or lead to incorrect conclusions.

Additionally, the lack of specific time and data stamps associated with the recorded data poses challenges in accounting for potential macroeconomic shocks and events that may influence poverty risk. Such factors are crucial for understanding the context in which the data was collected and may alter the results.

We also faced constraints related to the size of the dataset. The dataset did not reach the size we initially anticipated, as we aimed to combine both the test and training datasets. This limitation raises concerns about the robustness of our findings, particularly regarding the sample size across different groups and the potential bias in variable distributions.
Finally, our use of the linear model was constrained by computational time, limiting us to exploring only four predictors. As such, the depth of our analysis may have been restricted.

6 Conclusion

This study aimed to explore the feasibility of predicting poverty risk using a combination of socioeconomic, demographic and digital indicators. We investigated how these factors influence poverty risk across different age groups and created an interactive tool for policymakers and individuals to evaluate and mitigate poverty. By analysing the dataset, we built multiple models and identified the best-performing one to determine which independent and dependent variables serve as significant predictors of poverty.

Our analysis revealed several key findings, many of which align with previous studies. Age, employment status, educational level, financial activity, and phone technology usage were all significant predictors of poverty risk. Specifically, irregular employment and lower levels of education were associated with a higher probability of poverty. In contrast, increased engagement in financial activities and greater use of phone technology were linked to a reduced likelihood of poverty, indicating that digital access and financial literacy can help mitigate poverty risk.

These findings contribute to ongoing scientific research and societal relevance surrounding poverty by providing empirical evidence on how these factors shape poverty risk. Understanding these relationships is crucial for developing targeted policies to reduce poverty. For policymakers, the interactive tool offers a clear framework to prioritize interventions and strategies to address the most impactful factors. The ability to predict poverty risk through our interactive tool can be significant to empower policymakers take proactive steps towards poverty alleviation.

Nonetheless, several questions remain unanswered. Future research should explore the specific mechanisms through which these variables influence poverty risk, as well as identify additional overlooked factors. Examining developing economies could provide insight into how these relationships vary depending on local economic conditions or cultural contexts. Moreover, incorporating other potential predictors—such as healthcare access, time-specific data, and country identifiers—could further improve model accuracy and relevance.

Ultimately, this study provides valuable insights into the complex drivers of poverty and makes a meaningful contribution to the development of evidence-based, data-driven solutions for poverty reduction.

7 References

Afzal, M., Hersh, J., & Newhouse, D. (2015). *Building a better model: Variable selection to predict poverty in Pakistan and Sri Lanka* (World Bank Research Working Paper).

Cadena-Palacios, C. N., Araujo, I., Duque, R. A., & Benítez Soto, A. (2024). Inclusive targeting: A multistep validation of the poverty probability index for the identification of low-income students in developing countries. *Cogent Social Sciences, 10*(1). https://doi.org/10.1080/23311886.2024.2392024

Geda, A., de Jong, N., Mwabu, G., & Kimenyi, M. S. (2001). *Determinants of poverty in Kenya: A household level analysis* (Institute of Social Studies Working Paper Series No. 347). https://repub.eur.nl/pub/19095/wp347.pdf

Hohberg, M., Landau, K., Kneib, T., & Sturm, B. (2018). Vulnerability to poverty revisited: Flexible modeling and better predictive performance. *The Journal of Economic Inequality, 16*(4), 439–454. https://doi.org/10.1007/s10888-017-9374-6

Iqbal, N., & Awan, M. S. (2015). The impact of socioeconomic and demographic variables on poverty: A village study. *ResearchGate*. https://www.researchgate.net/publication/255659103_The_Impact_of_Socioeconomic_and_Demographic_Variables_on_Poverty_A_Village_Study

Maloma, I. (2016). The socioeconomic determinants of household poverty status in a low-income settlement in South Africa. *International Journal of Social Sciences and Humanity Studies, 8*(2), 122–134. https://dergipark.org.tr/en/download/article-file/257174

Mathiassen, A. (2013). Testing prediction performance of poverty models: Empirical evidence from Uganda. *Review of Income and Wealth, 59*(1), 91–112. https://doi.org/10.1111/roiw.12007

Pokhriyal, N., & Jacques, D. C. (2017). Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences, 114*(46), E9783–E9792. https://doi.org/10.1073/pnas.1700319114

Sekhampu, T. J. (2013). Determinants of poverty in a South African township. *Journal of Social Sciences, 34*(2), 145–153.

Verme, P. (2020). *Which model for poverty predictions?* (GLO Discussion Paper No. 468). Global Labor Organization. https://hdl.handle.net/10419/213811

Wimer, C., Nam, J., Garfinkel, I., Kaushal, N., Waldfogel, J., & Fox, L. (2020). Young adult poverty in historical perspective: The role of policy supports and early labor market experiences. *Social Science Research, 86*, 102390. https://doi.org/10.1016/j.ssresearch.2019.102390

Wu, Y., Naqvi, S. M. M. A., & Yasin, I. (2024). Dynamic relationship between social factors and poverty: A panel data analysis of 23 selected developing countries. *Journal of the Knowledge Economy, 15*, 19354–19386. https://doi.org/10.1007/s13132-024-01843-x
Yiu, J. (2019). *Poverty probability index & economic indicators*. Kaggle. https://www.kaggle.com/datasets/johnnyyiu/predicting-poverty?resource=download

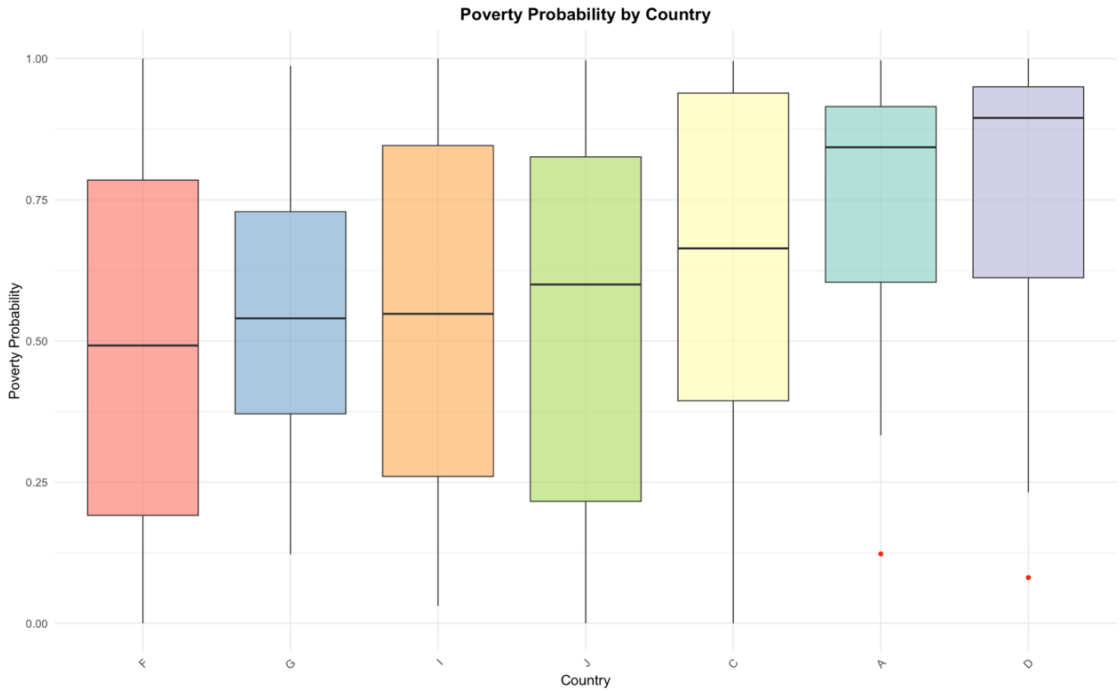
Zeller, M. (2013). Poverty status probability: A new approach to measuring poverty and the progress of the poor. *The Journal of Economic Inequality, 12*(4), 469–488. https://doi.org/10.1007/s10888-013-9264-5

Zhang, Y., & Wan, G. (2008). *Can we predict vulnerability to poverty?* (WIDER Research Paper No. 2008/82). The United Nations University World Institute for Development Economics Research (UNU-WIDER). https://hdl.handle.net/10419/45164

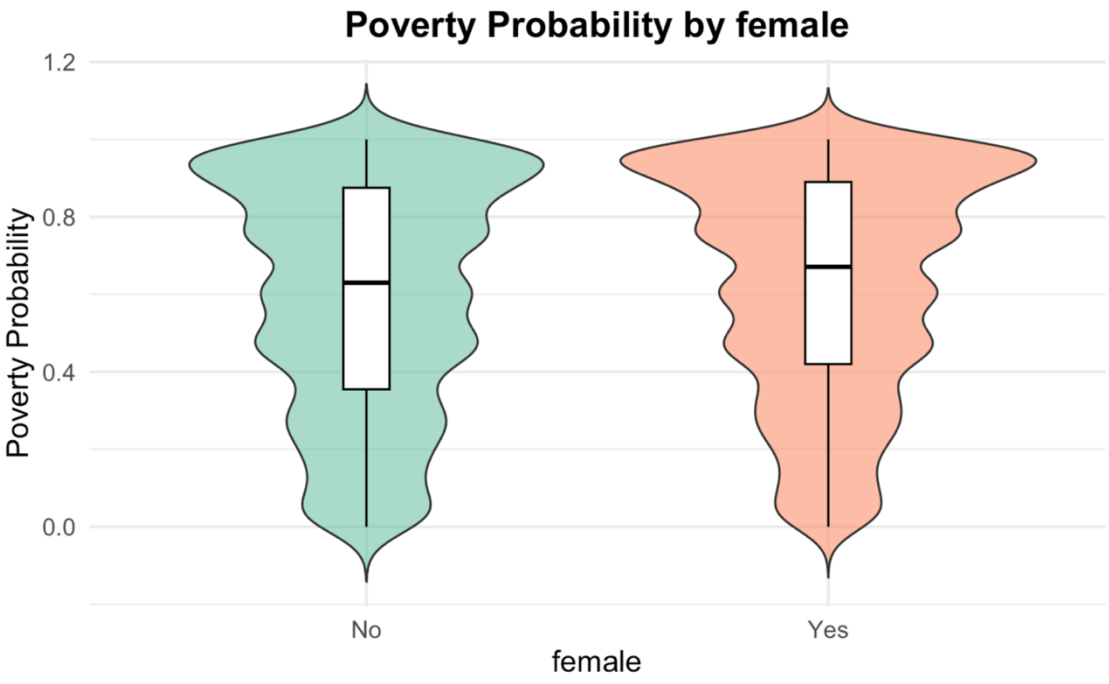
Zixi, H. (2021). Poverty prediction through machine learning. In *2021 2nd International Conference on E-Commerce and Internet Technology (ECIT)* (pp. 314–324). https://doi.org/10.1109/ECIT52743.2021.00073

8 Appendix

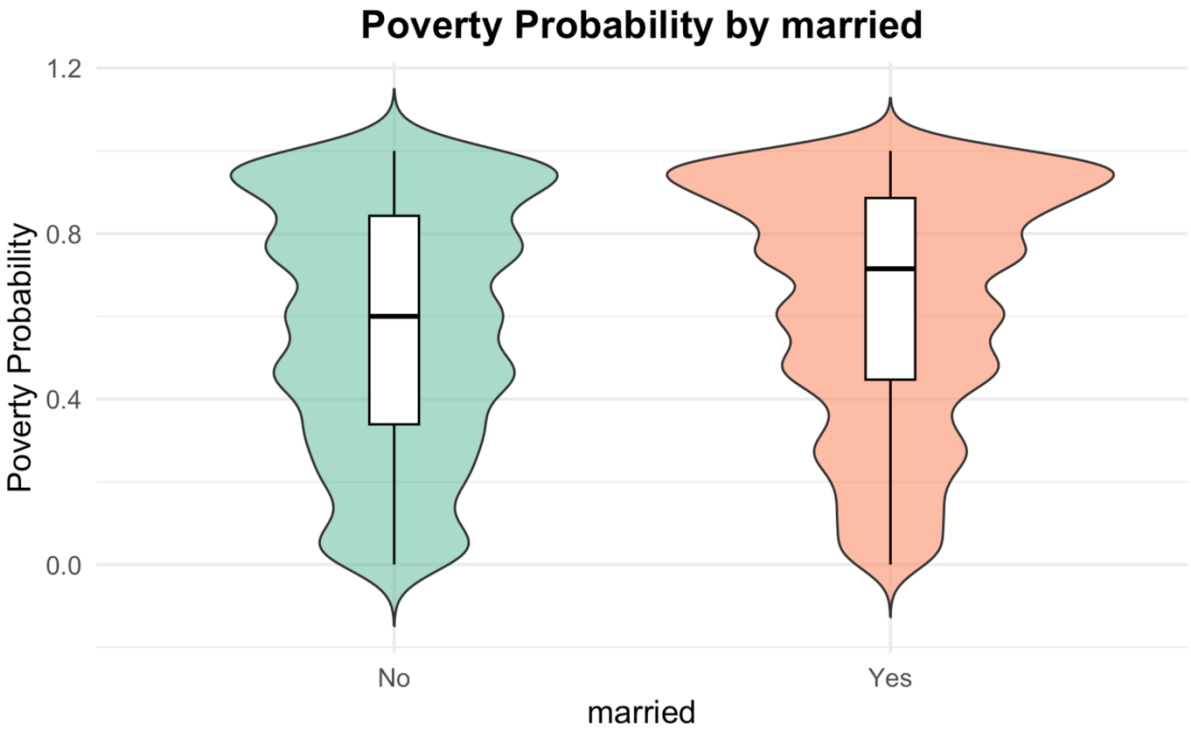
8.1 Visualisations



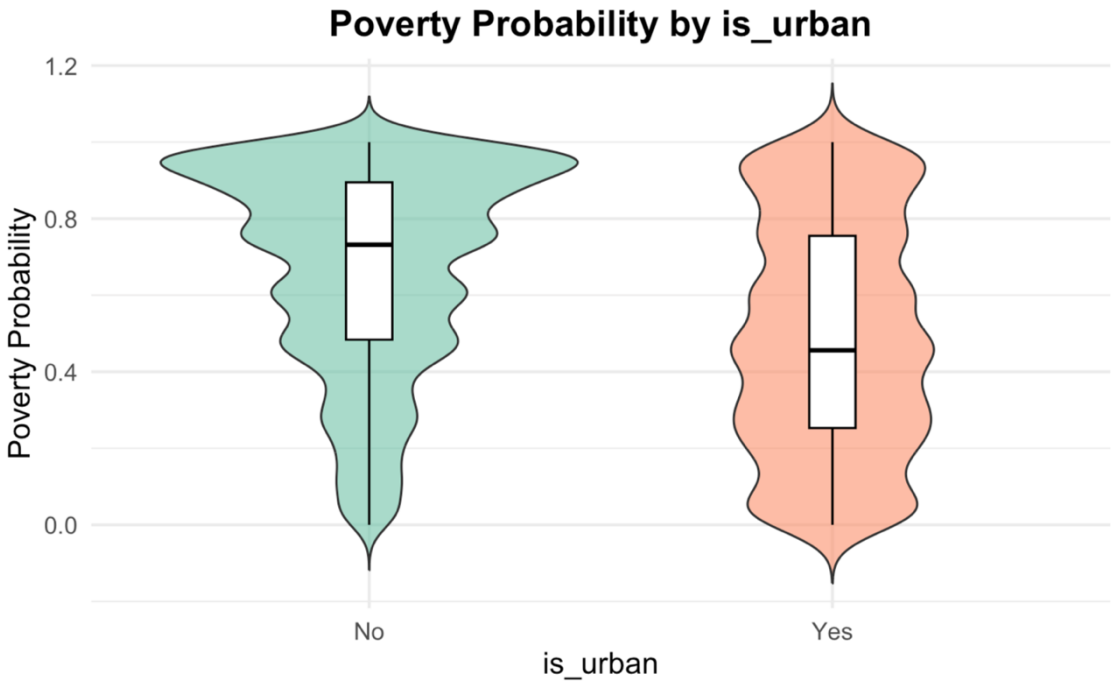
Appendix 1.1 Poverty Probability by Country



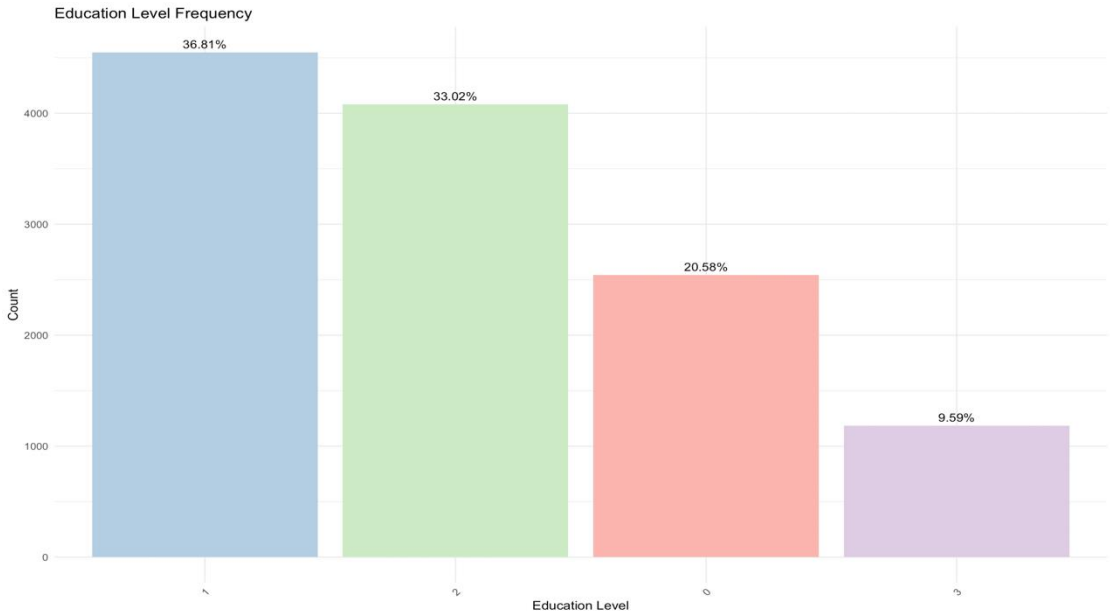
Appendix 1. 2 Poverty Probability by Gender



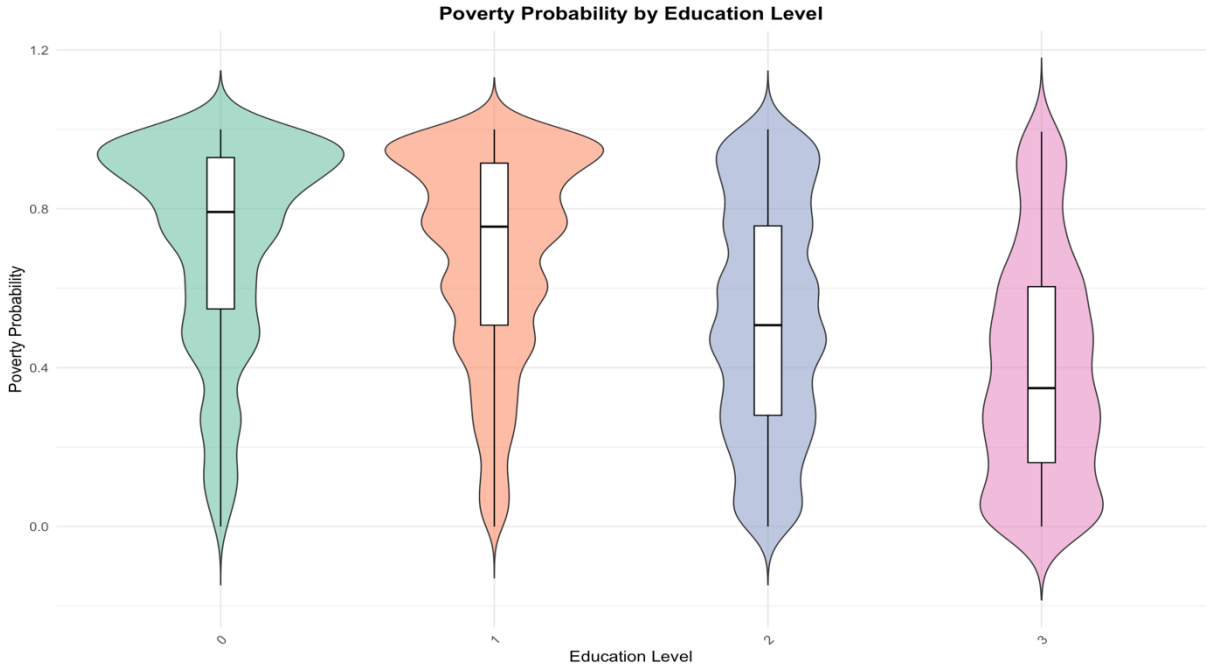
Appendix 1.3 Poverty Probability by Marital Status



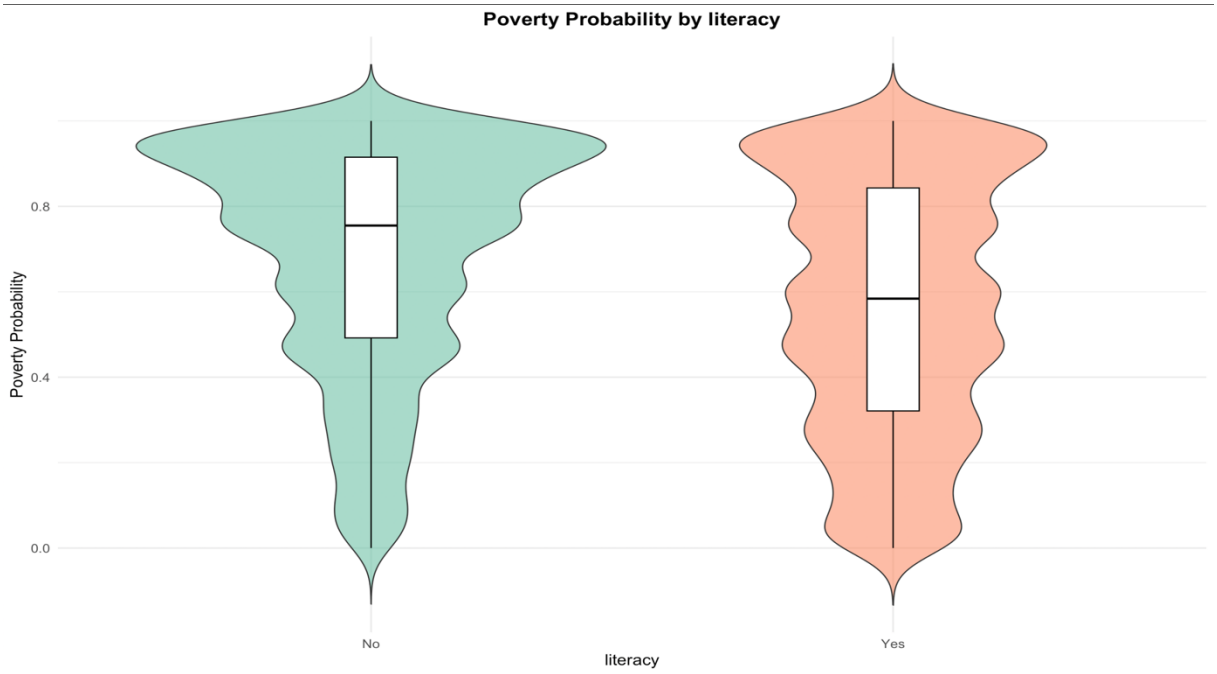
Appendix 1. 4 Poverty Probability by Living Area (Urban/ Rural)



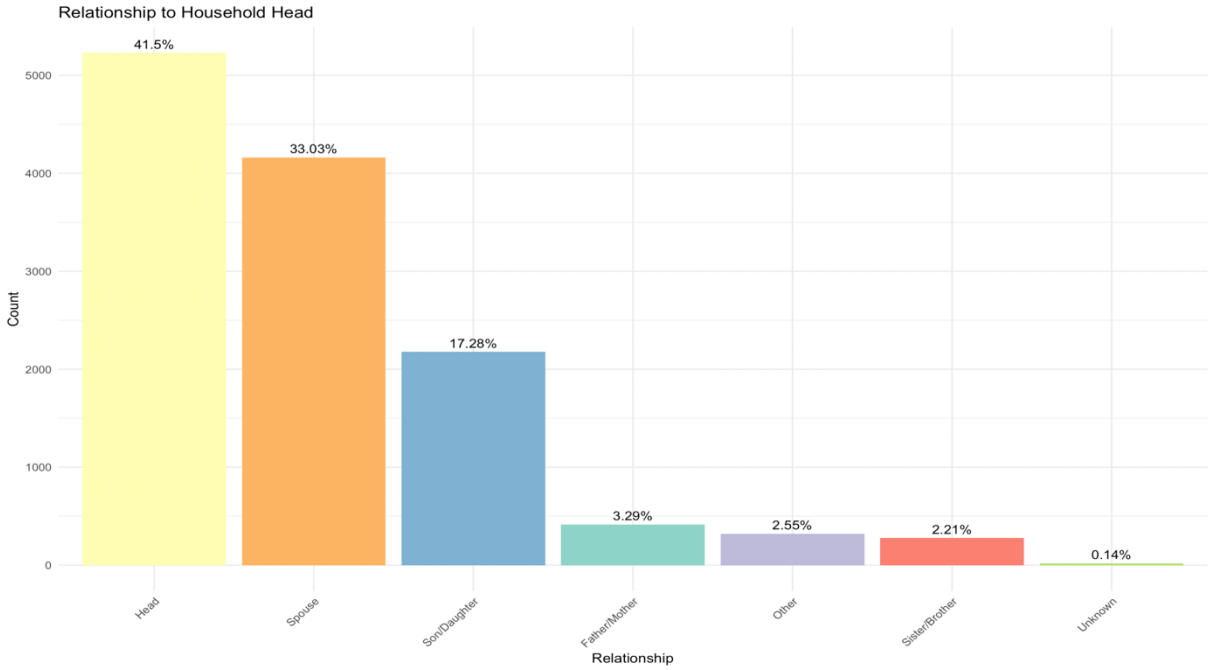
Appendix 1.5 Frequency Table for Education Level



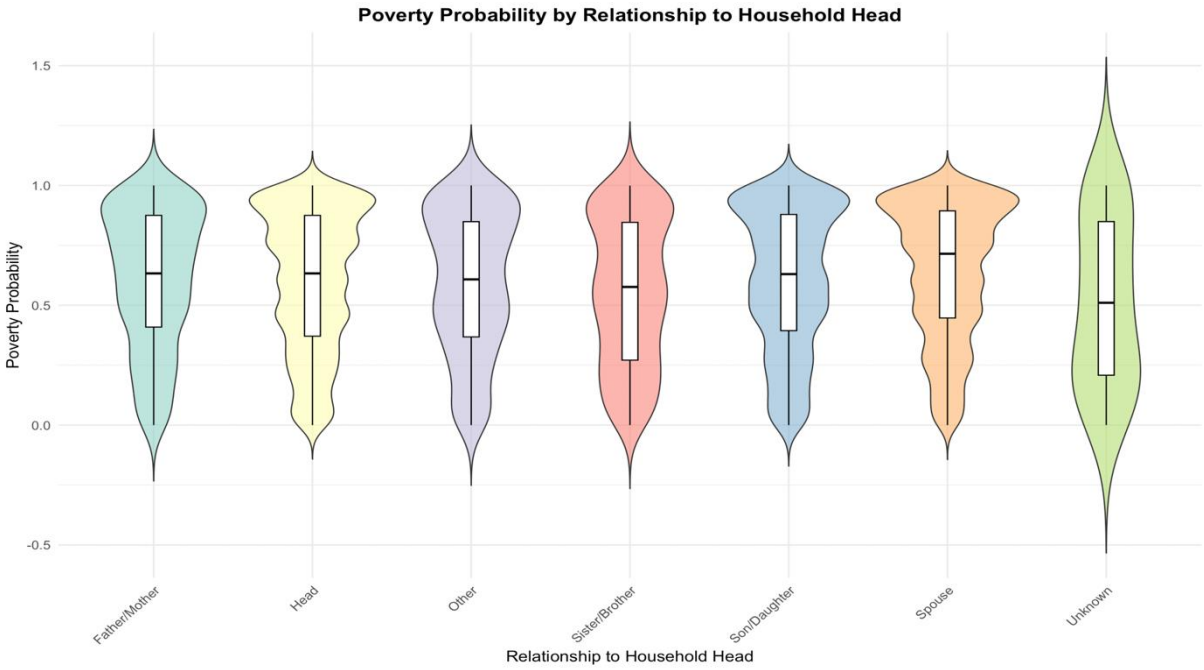
Appendix 1.6 Poverty Probability by Education Level



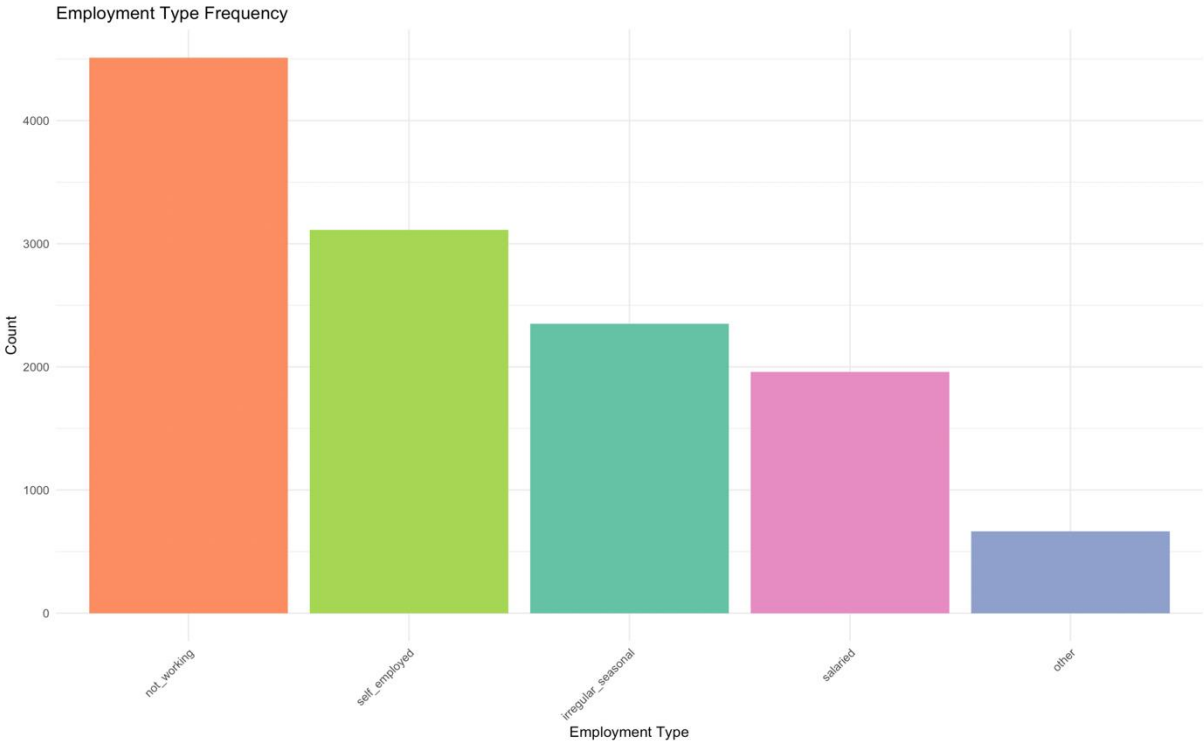
Appendix 1.7 Poverty Probability by Literacy



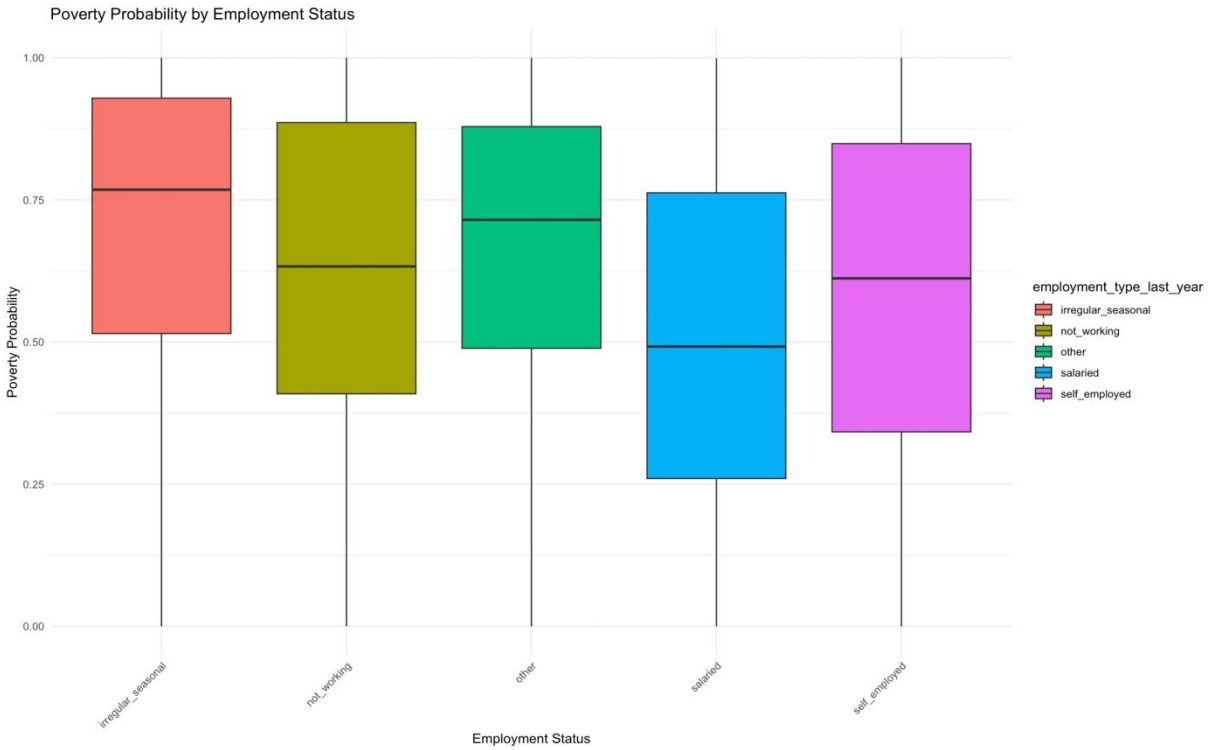
Appendix 1.8 Frequency Table for Relationship to Household Head



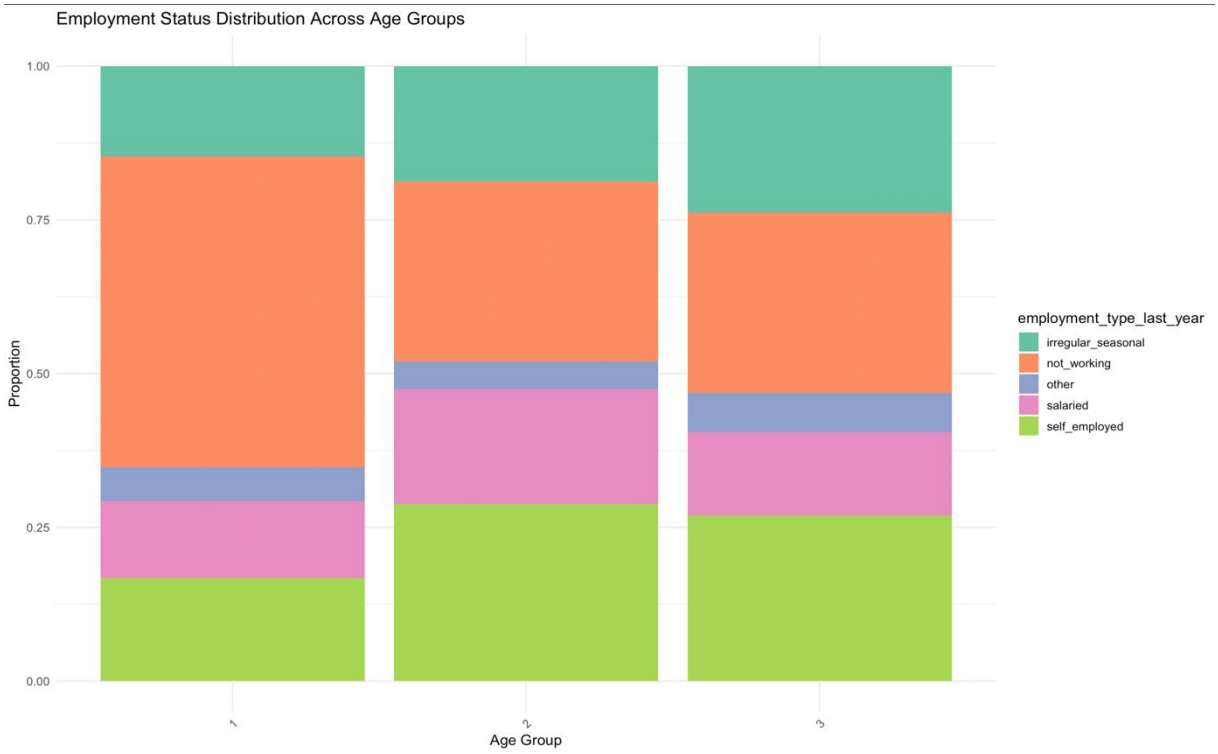
Appendix 1.9 Poverty Probability by Relationship to Household Head



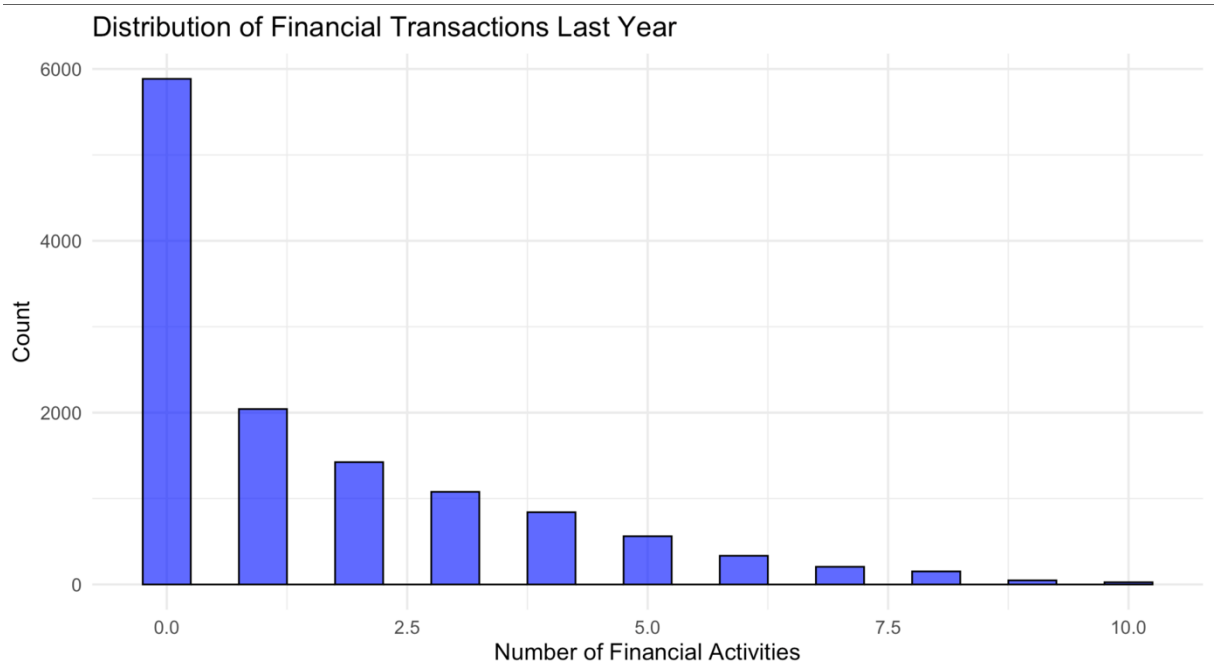
Appendix 1.10 Employment Type Frequency



Appendix 1.11 Poverty Probability by Employment Status



Appendix 1.12 Employment Status Distribution Across Age Groups



Appendix 1.13 Distribution of Financial Activities Last Year

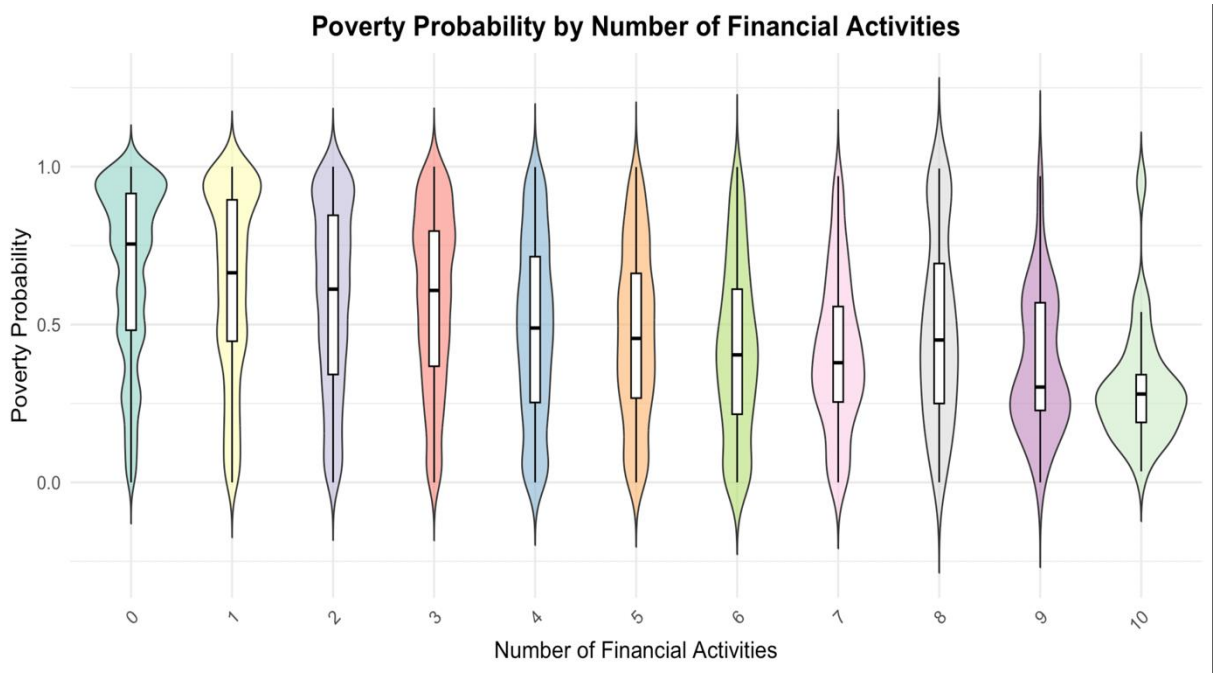
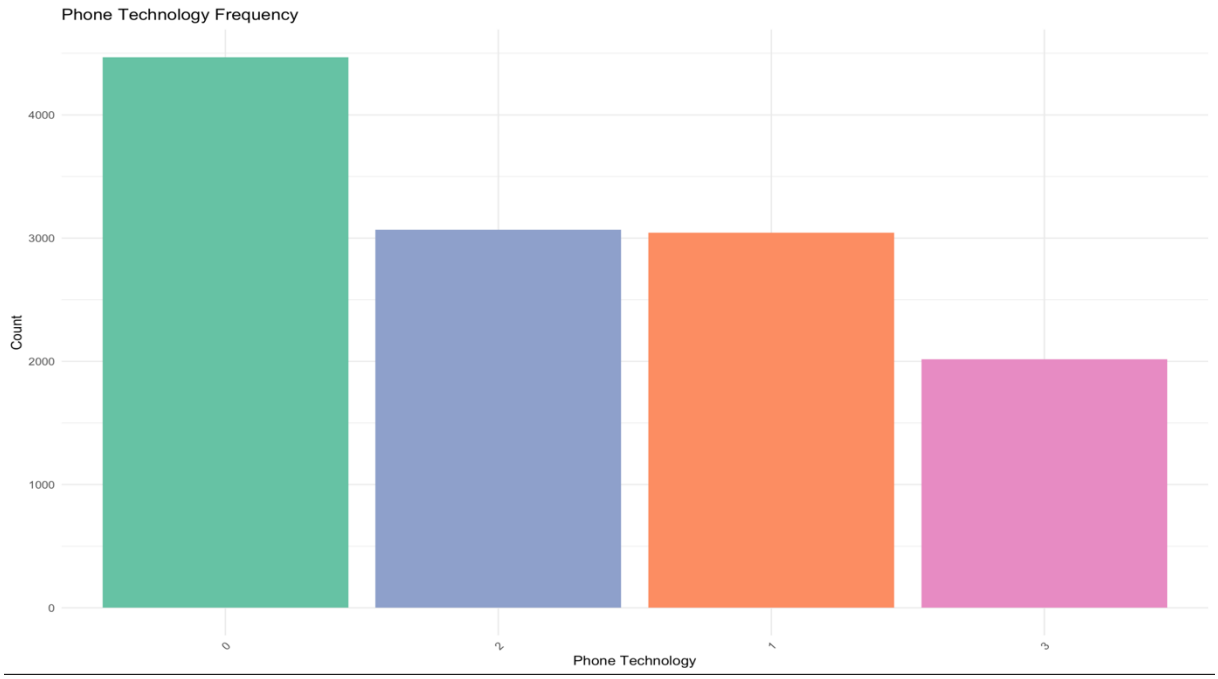
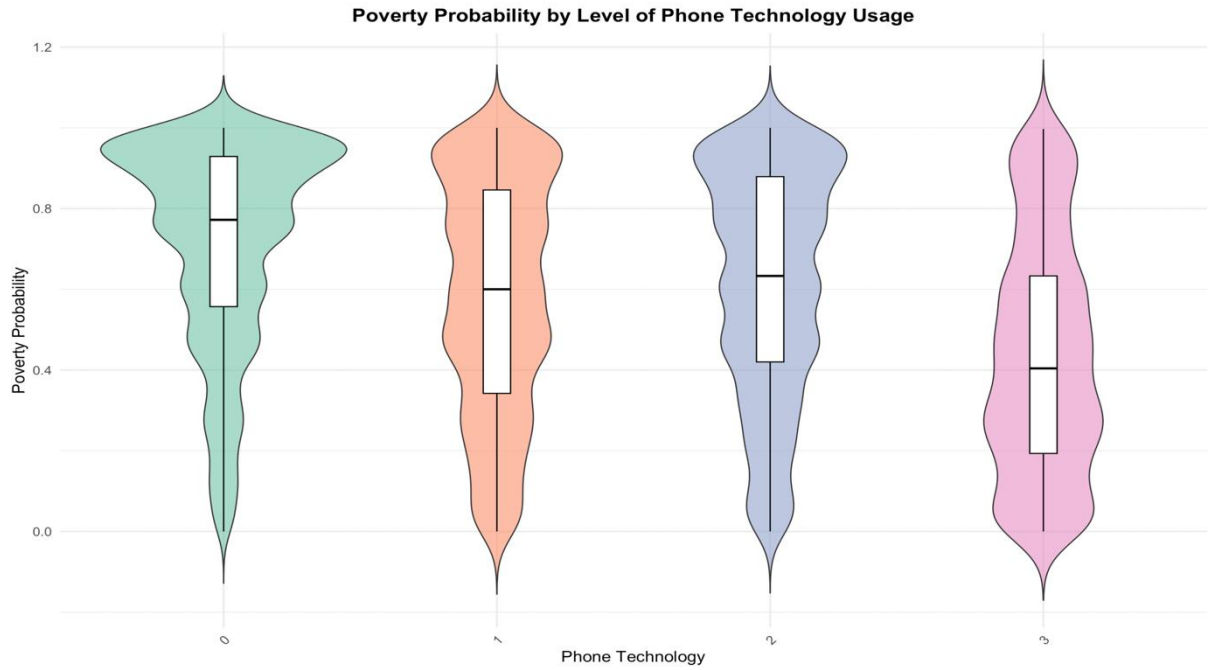


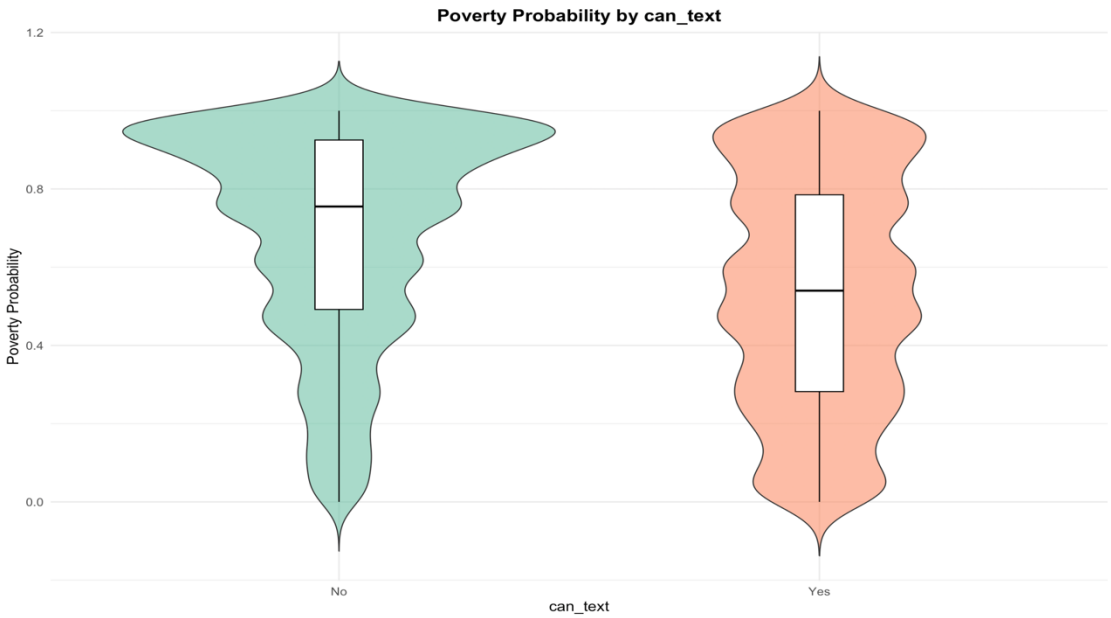
Figure 1.14 Poverty Probability by Number of Financial Activities



Appendix 1.15 Frequency Table for Phone Technology



Appendix 1.16 Poverty Probability by Level of Phone Technology



Appendix 1.17 Poverty Probability by ability to text