MORINGA SCHOOL

DATA SCIENCE PROJECT

Analysis of financial inclusion in East Africa

Authors:
Eric Cheruiyot
Felista Mueni
Antonia Mulinge
Joseph Leshakwet

Supervisors:
Samwel &
Veronica

Acknowledgements

We would like to thank our project supervisors Samwel & Veronica for their guidance and support throughout the topic selection and project proposal writing. Finally, we extend our appreciation to our friends who gave us lots of useful suggestions on high-level applications. ...

Chapter1

Introduction

Background of the Study

This project aims to analyze the level and scope of financial inclusion in East Africa's demography by using bank account ownership as the main indicator of financial inclusion. It will also provide insights to the demographic makeup and distribution of financial inclusion in Kenya, Rwanda, Tanzania, and Uganda.

Metric for Success

Financial Inclusion is one of the main obstacles to economic and human development in Africa. The unbanked who are a majority in East Africa are left behind as they lack access to credit and financial services. This only furthers the poverty rate and slows down economic growth.

Traditionally, access to bank accounts has been regarded as an indicator of financial inclusion. Despite the proliferation of mobile money in Africa and the growth of innovative fintech solutions, banks still play a pivotal role in facilitating access to financial services. Access to bank accounts enables households to save and facilitate payments while also helping businesses build up their credit-worthiness and improve their access to other financial services. Therefore, access to bank accounts is an essential contributor to long-term economic growth. For example, across Kenya, Rwanda, Tanzania, and Uganda only 9.1 million adults (or 13.9% of the adult population) have access to or use a commercial bank account.

Experimental Design

Below are the steps conducted in this analysis.

1. Load and preview the data.
2. Data Cleaning (check for and deal with outliers,checking for anomalies, messy column names, values and missing data)
3. Univariate Analysis
4. Bivariate Analysis
5. Multivariate Analysis
6. Implementing the Solution by performing the respective analysis i.e. factor analysis, principal component analysis, and discriminant analysis.
7. Challenging the Solution by providing insights on how you can make improvements.

Data Validation

The main data set contains demographic information and ownership of a bank account by individuals across East Africa. This data was extracted from various Finscope surveys ranging from 2016 to 2018.

The data available for this analysis is valid and useful towards answering the research question as itwill allow us to identify some of the key demographic factors that influence whether an individual
likely to have a bank account.

Understanding our data:

We have a data set that contains the following columns:

1. *country*:interviewee country
2. *year*: Year survey was done in.
3. *uniqueid*: Unique identifier for each interviewee
4. *has_a_bank_account* : Yes , No
5. *location_type*:Type of location: Rural, Urban
6. *cellphone_access*: If interviewee has access to a cellphone: Yes, No
7. *household_size:* Number of people living in one house
8. *age_of_respondent*: The age of the interviewee
9. *gender_of_respondent*: Gender of interviewee: Male, Female
10. *relationship_with_head*: The interviewee's relationship with the head of the house:Head of Household, Spouse, Child, Parent, Other relative, Other non-relatives, Don't know
11. *marital_status*: The martial status of the interviewee: Married/Living together, Divorced/Seperated, Widowed, Single/Never Married, Don't know
12. *education_level:* Highest level of education: No formal education, Primary education, Secondary
    education, Vocational/Specialised training, Tertiary education, Other/Don't know/RTA
13. *Type of job* :Interviewee *has* Farming and Fishing, Self employed, Formally employed Government, Formally employed Private, Informally employed, Remittance Dependent, Government Dependent, Other Income, No Income, Don't Know/Refuse to answer

We will start by performing EDA on our data set.

Load our data set and check out the first and last 5 rows of the data frame.

Our data set has 13 columns and 23524 rows/records.
Data Cleaning:

We start cleaning our data by checking for any missing data, outliers, messy column values and duplicates.

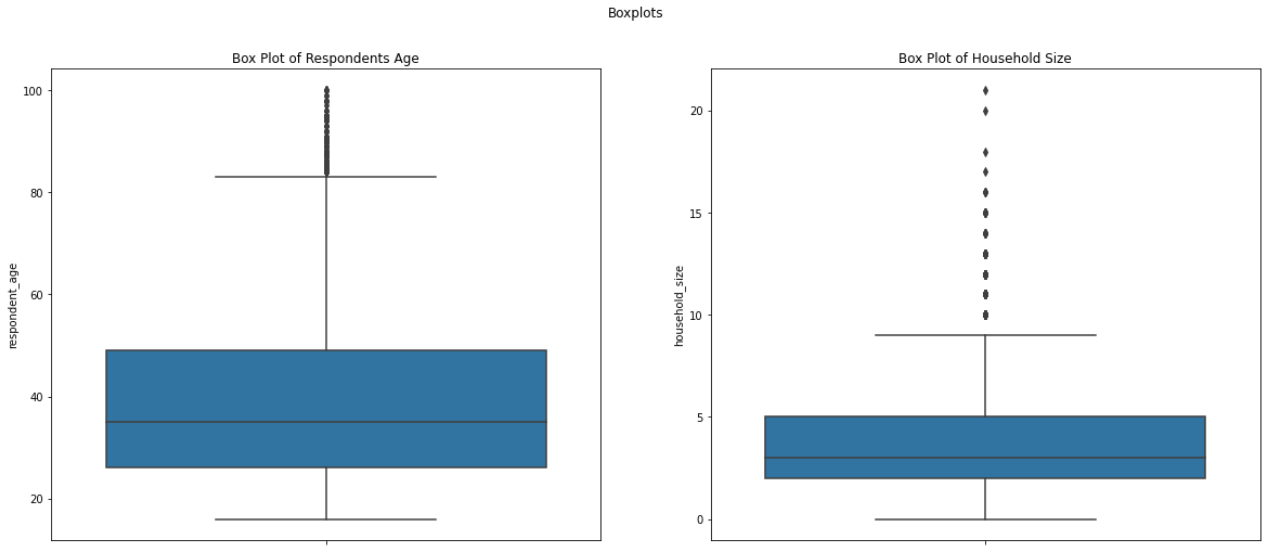
The data set contains 23525 rows and 13 columns after cleaning.

	country	year	uniqueid	Has a Bank account	Type of Location	Cell Phone Access	household_size	Respondent Age	gender_respon
0	Kenya	2018	uniqueid_1	Yes	Rural	Yes	3.0	24.0	Fe
1	Kenya	2018	uniqueid_2	No	Rural	No	5.0	70.0	Fe
2	Kenya	2018	uniqueid_3	Yes	Urban	Yes	5.0	26.0	
3	Kenya	2018	uniqueid_4	No	Rural	Yes	5.0	34.0	Fe
4	Kenya	2018	uniqueid_5	No	Urban	No	8.0	26.0	

	country	year	uniqueid	Has a Bank account	Type of Location	Cell Phone Access	household_size	Respondent Age	gender_o
23519	Uganda	2018	uniqueid_2113	No	Rural	Yes	4.0	48.0	
23520	Uganda	2018	uniqueid_2114	No	Rural	Yes	2.0	27.0	
23521	Uganda	2018	uniqueid_2115	No	Rural	Yes	5.0	27.0	
23522	Uganda	2018	uniqueid_2116	No	Urban	Yes	7.0	30.0	
23523	Uganda	2018	uniqueid_2117	No	Rural	Yes	10.0	20.0	

After cleaning the data we checked for outliers in our age and household sizes variables which are numerical and plotted the box plots as below.

We keep the outliers since it is possible to have people with ages greater than 80.
Also it is common for some families to leave together with extended family hence household size greater then 10 is possible



We then checked for anomalies in the cleaned data.The results show that 'respondent age' and 'household size' variables have 23 and 3 (respectively) recordsthat do not lie within the upper and lower bounds.

UNIVARIATE ANALYSIS.

The relationship between gender of the respondent and access to bank accounts
Count the number of respondents who can access bank accounts and those who can not access bank accounts.

3309 people had access to bank accounts while 20142 did not have access to bank accounts.

For the numerical data summary analysis we got the below output.

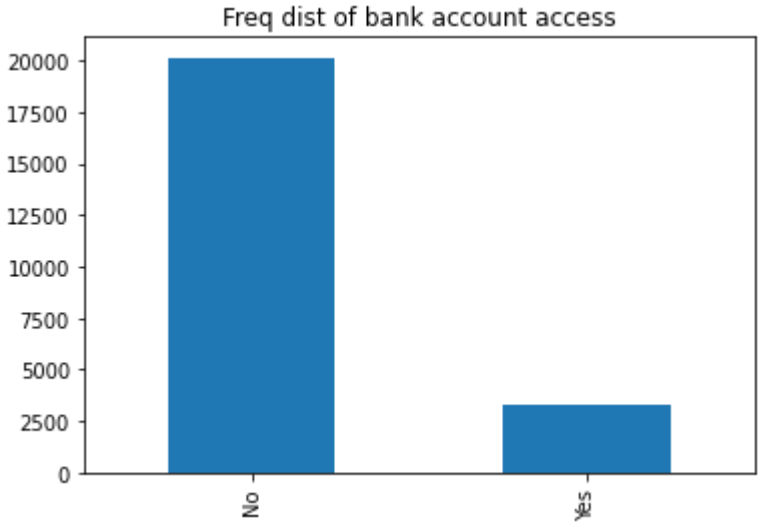
	country	year	uniqueid	has_a_bank_account	type_of_location	cell_phone_access	household_size	r
0	Kenya	2018	uniqueid_1	Yes	Rural	Yes	3.0	
1	Kenya	2018	uniqueid_2	No	Rural	No	5.0	
2	Kenya	2018	uniqueid_3	Yes	Urban	Yes	5.0	
3	Kenya	2018	uniqueid_4	No	Rural	Yes	5.0	
4	Kenya	2018	uniqueid_5	No	Urban	No	8.0	

	year	household_size	respondent_age
count	23447.000000	23447.000000	23447.000000
mean	2016.975690	3.686926	38.807417
std	0.848654	2.279248	16.516712
min	2016.000000	0.000000	16.000000
25%	2016.000000	2.000000	26.000000
50%	2017.000000	3.000000	35.000000
75%	2018.000000	5.000000	49.000000
max	2018.000000	21.000000	100.000000

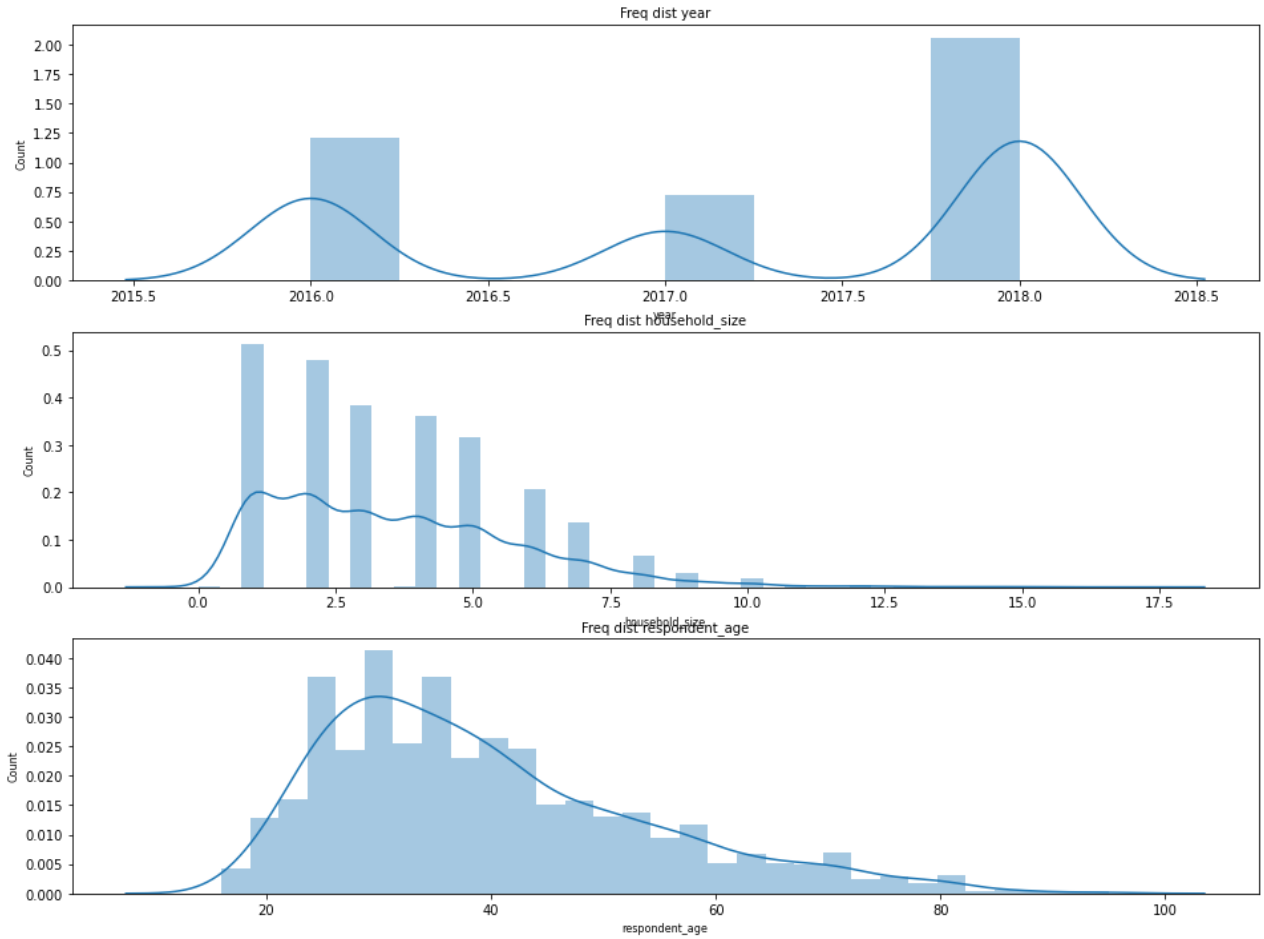
Summary:

The data set contains 23447 rows and 9 columns after dealing with the missing values.There are 3309 people who canaccess bank accounts in the data set, which is where I'll focus my analysis.

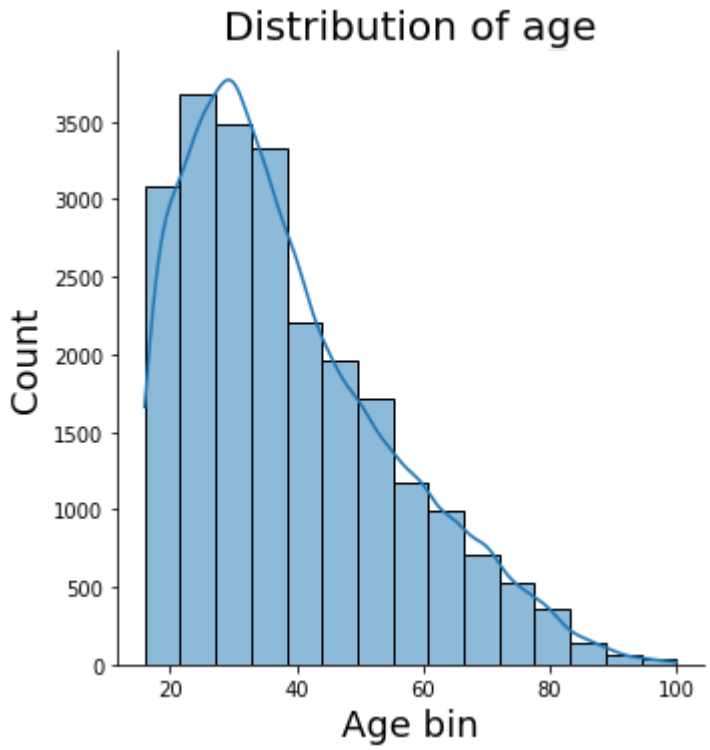
Has access to bank account column is a categorical variable
For categorical columns we plot histograms, we use the value_count() and plot.bar() functions.



We then plot the univariate distribution of the numerical columns which contains the histogram.



We can see that leaving the Year column every other column is skewed to the left which indicates most of the values lie in the lower range values.The year attribute shows normal distribution.# Analysing our respondents age for distribution density.

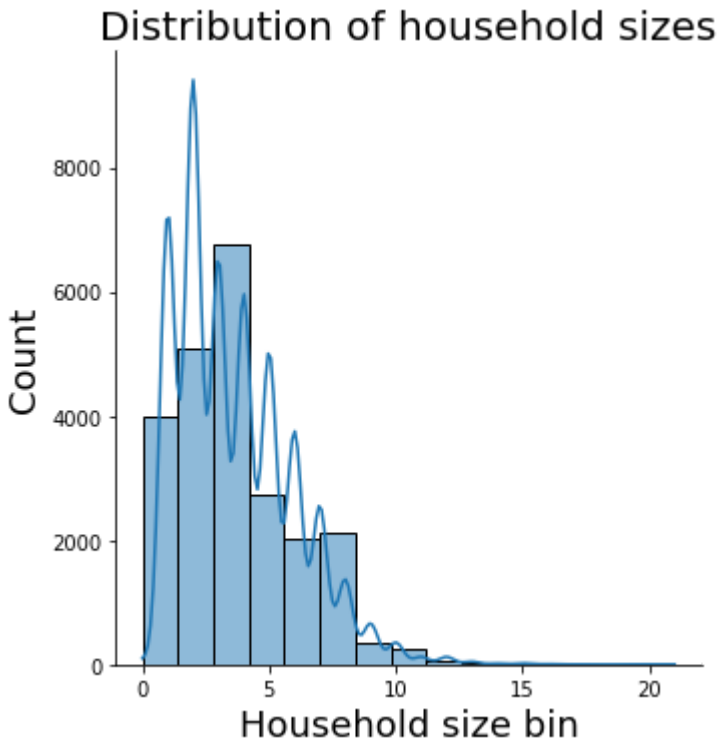


The descriptive statistic of the data

count	23447.000000
mean	38.807417
std	16.516712
min	16.000000
25%	26.000000
50%	35.000000
75%	49.000000
max	100.000000

From the table the minimum age of the respondent is sixteen years while the maximum age of the respondents is 100 years. The average age of the respondent is 38 years.

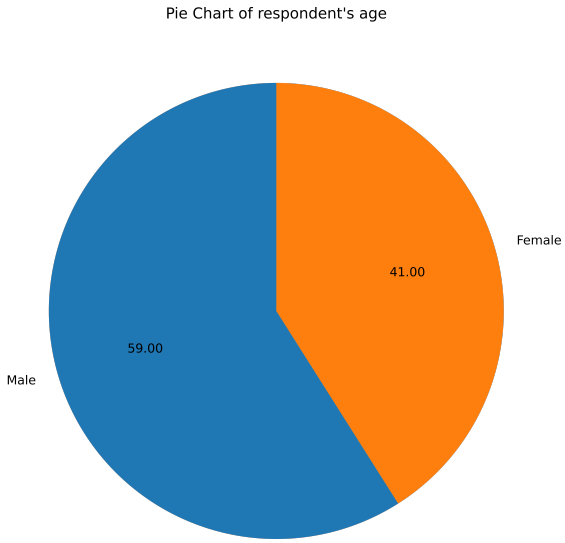
We analysed the distribution of household sizes by plotting a distribution histogram and did descriptive data analysis of the household sizes data.

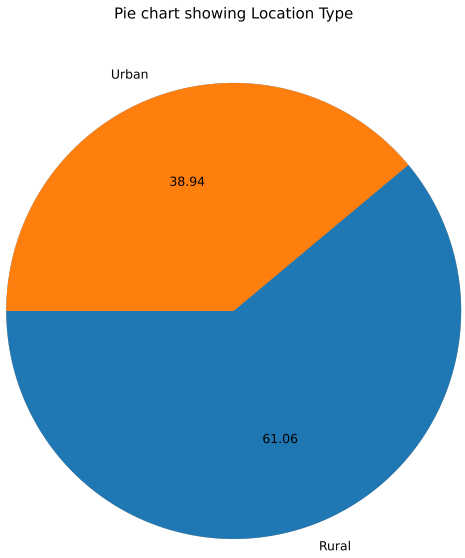



```
count    23447.000000
mean         3.686926
std          2.279248
min          0.000000
25%          2.000000
50%          3.000000
75%          5.000000
max         21.000000
```

The average household size of the data is 3 people per household the maximum household size is 21 people in the household.

We then plotted pie charts of the ages and locations of the respondents which is categorical data.





Measures of Dispersion

We did measures of dispersion analysis on the age and household sizes of the respondents and got the following as output:

The Respondent's age standard deviation is 16.516711802713548 The Respondent's age variance is 272.801768773897

The range of the Respondent age is : 84.0

The Interquartile Range for Respondent age variable is : 23.0

The skewness of age variable is : 0.8418778832863799

The kurtosis of age variable is : 0.10346047874300401

Skewness is positive meaning that data is skewed right. Since Kurtosis is greater than zero , we can say data is heavy tailed hence there's presence of outliers

```
The Respondent's Household size standard deviation is  2.279247718307969
The Respondent's Household size variance is  5.1949701614120825
The range of the Respondent Household size is :  21.0
The Interquartile Range for Respondent Household size variable is :  3.0
The skewness of Household size variable is :  0.9752673119630328
The kurtosis of Household size variable is :  1.1580661329361344
```

Skewness is positive meaning that data is skewed right. Since Kurtosis is greater than zero , we can say data is heavy tailed hence there's presence of outliers

DEDUCTIONS FROM UNIVARIATE ANALYSIS

Numerical variables:

Age distribution was skewed to the right. The mean age was 38 , median was 35 , mode was 30.This shows that most of our respondents were in their thirties.

The Household size distribution was skewed to the right. Mean was 3 , median was 3 and mode was2.

Categorical variables:

. From the dataset, we can say that there were more males than females interviewed.

. Also, most respondents reside in the rural areas compared to urban areas.

. Most of the respondents were heads of the household and were married or living together with their spouses.

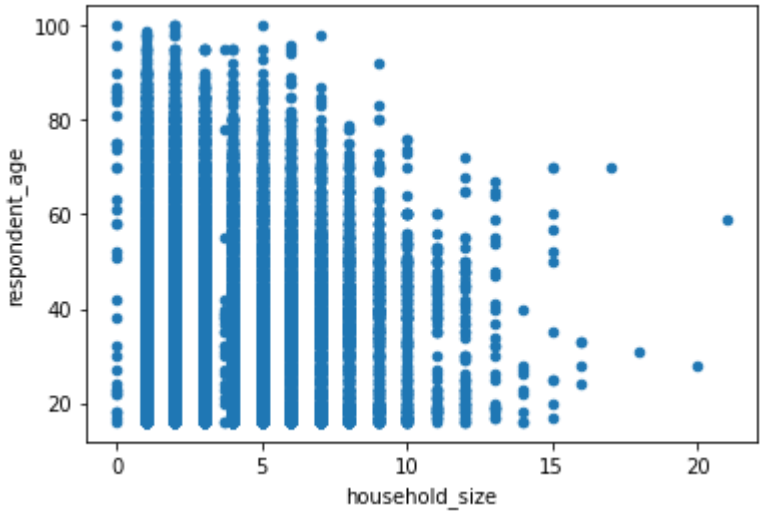
. Most of them had access to cellphones and were self employed. Most of the respondents did not have bank accounts.

BIVARIATE ANALYSIS

Numerical-Numerical

For numerical-numerical we do scatter plot and linear correlation where scatter plot

The bivariate distribution plots help us to study the relationship between two variables



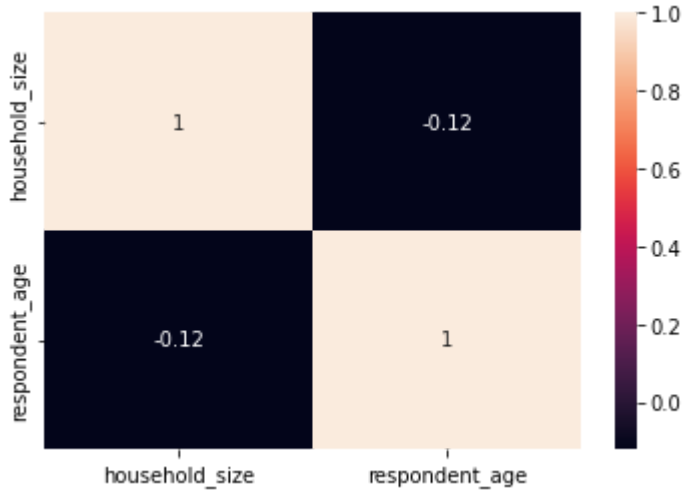
The scatter plot show that as the house hold size increases,the respondent age decreases.
We then analysed linear correlation to show the strength of the relationship between the two variables.

```
correlation = -0.11980831235152938
```

There is Weak negative correlation between house hold size and respondent age.This means if we increase one variable the the 2nd variable decreases with the same amount

Plotting the Pearson correlation coefficient among numeric variables

We can see that the two variables below are not correlated since the correlation coefficients are close to 0.



Categorical - Categorical

We compared various categorical variables with each other.
To begin we checked how many individuals have bank accounts, based on their gender and type of location

type_of_location		Rural	Urban	All
gender_of_respondent	has_a_bank_account			
Female	No	7797	4554	12351
	Yes	724	758	1482
Male	No	4853	2938	7791
	Yes	943	880	1823
All		14317	9130	23447

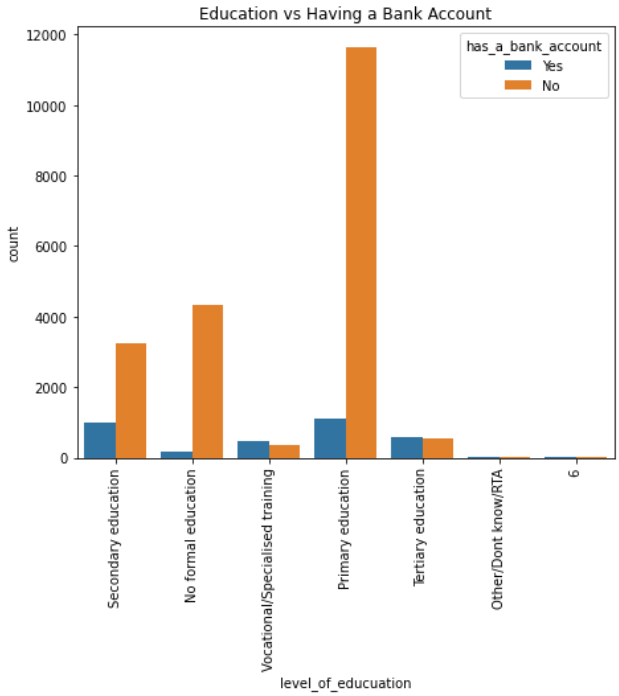
We see that generally, more men have bank accounts compared to women.It also shows that females who live in urban areas are more likely to have bank accounts relative to those who live in rural areas. The opposite is true for males.

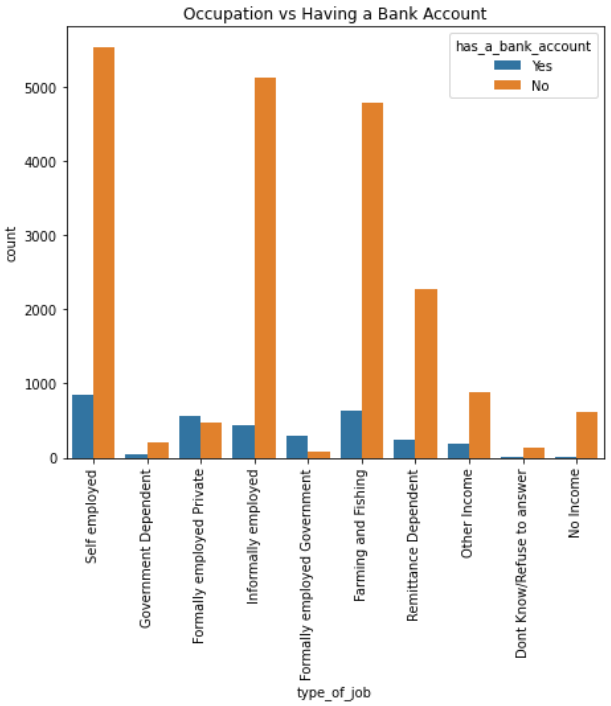
We also checked how many individuals have bank accounts, based on their marital status and also their phone ownership status.

marital_status		Divorced/Seperated	Dont know	Married/Living together	Single/Never Married	Widow
cell_phone_access	has_a_bank_account					
No	No	609	8	1915	2328	10
	Yes	7	0	41	30	
Yes	No	1221	24	6950	4637	13
	Yes	225	4	1823	931	2
All		2062	36	10729	7926	26

We see that generally, individuals who have cell phones are likely to have have bank accounts compared to those who don't have cell phones.

Comparing education level and occupation vs having a bank account.





It is seen that level of education mattered a lot as far as having bank account is concerned.

Those who had primary and secondary education have more access to bank accounts.

The self employed also had more access to bank accounts as compared to other types of job.

We changed some categorical attributes to numeric so that we can get a summary of all correlations to whether a person has a bank account or not.

1. Correlation between type of location with bank account status is -0.08824344467617874

2. Correlation between cell phone access with bank account status is 0.20943545716751724

3. Correlation between household size with bank account status is -0.022222336818058

4.Correlation between household size with bank account status is 0.11658534225280977

Correlation coefficients between -.20 and .20 are generally considered to be weak and therefore type_of_location, cell phone access, house hold size and gender of respondents have weak correlation with access to bank accounts.

Bivariate analysis findings
Numerical:

Younger people had bank accounts compared to the elder people Household with less people had bank accounts.

There was a very small correlation between the household size and one owning a bank account. There was a very small correlation between the household size and one owning a bank account.

Categorical:

Respondents who were head of their households, Most of them had no bank accounts. Among those with bank accounts, they were also the most. This may be because they were the majority in that demograph.Respondents who were married were the majority among those with bank accounts and also those without.

We noted that a majority of respondents with only a Primary level of education had no bank accounts. However respondents with any form of education, e.g secondary, vocational , tertiary, most of them had bank accounts.

Most respondents had access to cell phones and yet did not have bank accounts.

Most respondents who were female did not have bank accounts.

Problems and Recommendations

Our data is highly imbalanced. It leans more towards certain demographics than others.

Also our data is collected from different years for different countries hence may result to some minor inaccuracies.

MULTIVARIATE ANALYSIS

We will use Principal Component Analysis (PCA) to select the most important features in the data set that tell us the maximum amount of information about the data set

Preprocessing:

The first preprocessing step is to divide the data set into a feature set and corresponding labels. We can store the feature sets into the X variable and the series of corresponding labels in to the y variable.We then Split the data set into the Training set and Test set.
Lets define the size of the test data as 20% of entire data set.
We then performed standard scalar normalization to normalize our feature sets and applied PCA having not specified the number of components in the constructor.We then checked the variance caused by each of the principal components using the explained variance ratio.

array([0.22974683, 0.18793178, 0.17013781, 0.11555847, 0.09416687,0.08051338, 0.07412638, 0.04781848])

The result above shows that the first principal component is responsible for 22.96% variance.The 2nd principal component causes 18.89% variance in the data set.summation of 1st and 2nd principle components ie (22.97 + 18.79) gives 41.76% .Therefore 41.76% of the classification information contained in the feature set is captured by the first two principal components.

We then used 2 Principal Component to train our algorithm in making predictions using random forest classification and got the output below.

[[4027 0]
[663 0]]

Accuracy is 0.85863539445629

With two features, the random forest algorithm is able to correctly predict an 85.86% accuracy.

We then 1 principal component to train the algorithm and make predictions using random forest classification and got the output below.

[[4027 0]
[663 0]]

Accuracy 0.85863539445629

With only 1 feature, the random forest algorithm is able to correctly predict an 85.86% accuracy.

Discriminant Analysis

We evaluated the “factorability” of our data set using Bartlett’s test. If Bartlett’s test turns out to be statistically insignificant, then we cannot use a factor analysis.
The output:

(27040.367226067192, 0.0)

Here the Bartlett ‘s test shows that the p-value is 0. The test was statistically insignificant, indicating that the observed correlation matrix is not an identity matrix.

Next we checked the Kaiser-Meyer-Olkin (KMO) Test. This will help us to determine the adequacy for each observed variable and for the complete model.
The output:

0.5363302951130932

The overall KMO for our data is 0.54, which is also considered inadequate since kmo of less than 0.6 is considered inadequate and therefore we cannot continue with planned factor analysis.

The overal kmo for our data is 0.54 which is inadequate

Conclusions and Recommendations

People with mobile phone access are very likely to have bank accounts.Banks can therefore utilize mobile banking capabilities to reach more people.There is no big difference between people living in rural vs urban areas as far as having a bank account is concerned. Individuals with Formal and Government jobs are more likely to have bank accounts.

Financial Institutions should introduce Digital Financial solutions that may bring about inclusion in East Africa. Too much focus on brick and mortar systems isolates the upcoming generation that is more tech-savvy and educated.There is a growing population of unbanked women who stand a chance to benefit given that financial institutions amend or provide digital solutions. Such institutions are likely to benefit by gaining more customers and also enhancing the lives of people.