**GROUP NAME: Brainy Badgers**

**GROUP LEAD: Kevin Kilonzo**

**GROUP MEMBERS:**

1. Patrick Kiplimo
2. Brenda Chepkoech
3. Stacy Keago
4. Kevin Kilonzo
5. Brandon Bosire

## **1.Business Understanding**

### **Business Overview**

Financial inclusion remains one of the main obstacles to economic and human development in Africa. For example, across Kenya, Rwanda, Tanzania, and Uganda only 9.1 million adults (or 14% of adults) have access to or use a commercial bank account. In the recent past, there has been a significant growth of innovative fintech solutions like MPESA services.

Our research focuses on how mobile money transactions can lead to greater financial inclusion for people who don't have access to banks. Greater financial inclusion leads to greater access to credit services, and easier tax base to achieve revenues, better access to savings and fintech services.

### **Business Objective**

To determine the number of people who have access to mobile phones and don't have access to the banks so as to advise the fintech companies on the areas or markets that they should target.

### **Business Success Criteria**

To compile a report on the regions which need improved infrastructure in terms of mobile phone services and which groups to be targeted for effective marketing for increased access to fintech services.

**Assessing the Situation**

1. **Resource Inventory**
   1. Datasets:
      1. Train
      2. Variable definition
   2. Software( Github, Google Collaboratory, SQLite, JIRA)
2. **Assumptions**
   1. The data provided is correct and up to date
3. **Constraints -**There are no constraints

### **Data Mining Goals**

Our data mining goals for this project are as follows:

* To determine the number of people with access to bank accounts and how this varies in urban and rural areas, between the job type.
* To determine the country with the most mobile phones and the highest number of bank accounts.
* To determine the people to be targeted with mobile banking intergration
* To determine the job type to be targeted with fintech marketing in the region for people withTo determine the influence of mobile phone access and location to job type a phone but no bank account
* To determine the influence of mobile phone access and location to job type
* To determine which age has the most access to mobile phones
* To detemine the Age group with mobile phone but no bank account
* Determine which age has the highest access to mobile phones and bank accounts.
* To determine the percentage of level of access to mobile phones for people in the region per country
* To determine the percentage of level of access to mobile phones for people in the region.
* To determine the percentage of level of access to mobile phones for people in urban areas and rural areas

**Data Mining Success Criteria**

Our success criteria will be measured by the following criteria;

* We target the groups that can be marketed for greater market share and increased accessibility to fintech services.

**2. Data Understanding**

### **Data Understanding Overview**

For this project, we are using the dataset from Zindi website. These datasets are;

* Train - This dataset contains demographic information and what financial services are used by a sample of the East African population.
* Variable Definition- This dataset contains the description of the Train dataset

### **Data Description**

We have two datasets available for this project. A detailed description of the datasets is provided as follows:

**Train -** This dataset contains demographic information and what financial services are used by a sample of the East African population. It has 13 columns and 23525 rows. This data was extracted from various Finscope surveys ranging from 2016 to 2018.

**Variable Definition-** This dataset contains the description of the Train dataset and the column definitions.The attributes are as follows:

* Country - (Country interviewee is in.)
* year - (Year survey was done in.)
* unique id - (Unique identifier for each interviewee)
* location type - (Type of location: Rural, Urban)
* cellphone access - (If interviewee has access to a cellphone: Yes, No)
* household size - (Number of people living in one house)
* Age of respondent - (The age of the interviewee)
* gender of respondent - (Gender of interviewee: Male, Female)
* relationship with head - (The interviewees' relationship with the head of the house:Head of Household, Spouse, Child, Parent, Other relative, Other non-relatives, Dont know)
* maritial status - (The martial status of the interviewee: Married/Living together, Divorced/Seperated, Widowed, Single/Never Married, Don't know)
* education level - (Highest level of education: No formal education, Primary education, Secondary education, Vocational/Specialised training, Tertiary education, Other/Dont know/RTA)
* job type - (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, Dont Know/Refuse to answer)

### **Verifying Data Quality**

None of the two datasets had any missing values. There were also no known data errors in the datasets.

## **3.Data Preparation**

These are the steps followed in preparing the data

#### **Loading Data**

Loaded the datasets from the CSV and then created a colab notebook from them.

#### **Cleaning Data**

Prior to the analysis of the dataset, cleaning will be undertaken. Primarily this will involve the removal of duplicates, removal of the ‘uniqueid’ column, and the conversion of incorrect column data types as required.

The next step will be the creation of new columns from the breaking of existing columns. From the relationship with the head column, we will create a new column called head\_of\_household, this will state yes if the respondent is the head of the household.

The next column to be broken will be the marital status column creating a new column called the living\_condition which will identify the state of the respondent whether they live with their spouses or not.

To then ensure the validity of the data we will then identify duplications in the dataset, dropping all duplicates from the data.

#### **Deriving New Attributes**

After cleaning our dataset, we created a new column named “head\_of\_household,”. This column was populated by converting the ‘relationship with head’ column into two having the(/) as a delimiter. This caused all the values that came after the delimiter to take up the place of the newly created column.

## **4. Analysis**

During our analysis, we were able to identify, through the returned results for the following research objectives ;

**To determine the number of people with access to bank accounts and how this varies in urban and rural areas**

From our analysis, we found that the higher number of bank accounts were in the rural areas,these were 1,671 accounts. While the urban areas had bank accounts reaching 1,641.From this it is very clear that there it is advisable to show no bais among the two.

**To determine the numbers of people with access to bank accounts and how this varies with job type**

From the visualisation below, it can be noted the relationship between bank account availability and job types.

**To determine the country with the most mobile phones and the highest number of bank accounts.**

To do this we broke the task into two parts.The first identified the country with the most number of cellphones, and the second identified the country with the most bank accounts.From these it was concluded that the two categories did not come from the same county:

**i) To determine the country with the most bank accounts**

Kenya was identified as being the country with the highest number of bank accounts. This being 1,521.

**ii)To determine the country with the most mobile phones.**

Rwanda was concluded as being the country with the highest number of cellphone accessing person.This number

was 7,249 people.

**To determine the people to be targeted with mobile banking integration**

These were identified first by a calculation of the number of persons with jobs as well as bank accounts but lacked cellphone access.The total of these is 3,197 persons.The jobs to be targeted constitute of this total as follows:

Cellphone count Job type.

831 Self employed

605 Farming and Fishing

568 Formally employed Private

435 Informally employed

296 Formally employed Government

227 Remittance Dependent

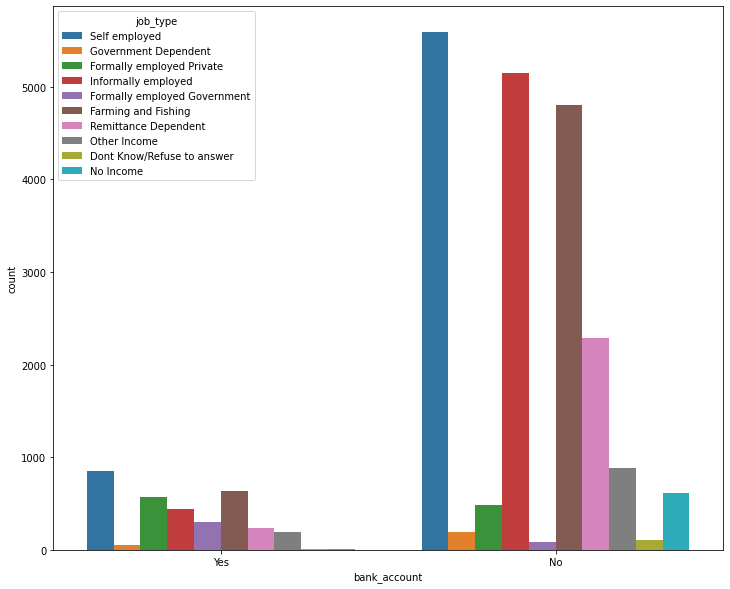
186 Other Income

35 Government Dependent

14 Don't Know/Refuse to answer

**To determine the job type to be targeted with fintech marketing in the region for people with a phone but no bank account**

This task was to find out how the job types count for those with phones corresponded with number of bank accounts for each respective job count.This was done through visualization of bank account access (whether there was any or whether their wasn’t ) and the job types count.



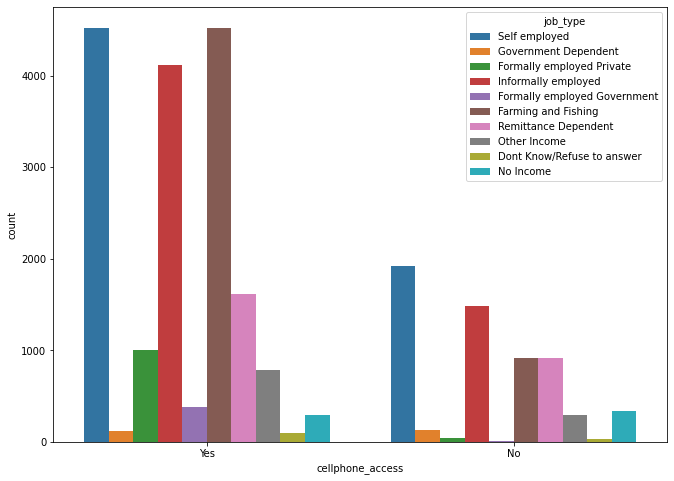
From the above, it has been noted that the largest portion of job types counts belonged to person without any available bank accounts .This means that there is a large market for persons without bank accounts with income sources who can be targeted for fintech services.

**To determine the influence of mobile phone access and location to job type**

Through visualization of the dataset, there was a trend identification task in relation to the cellphone access ,respondent’s location type (whether in an urban or a rural area) and any perceived effect on the job type counts.

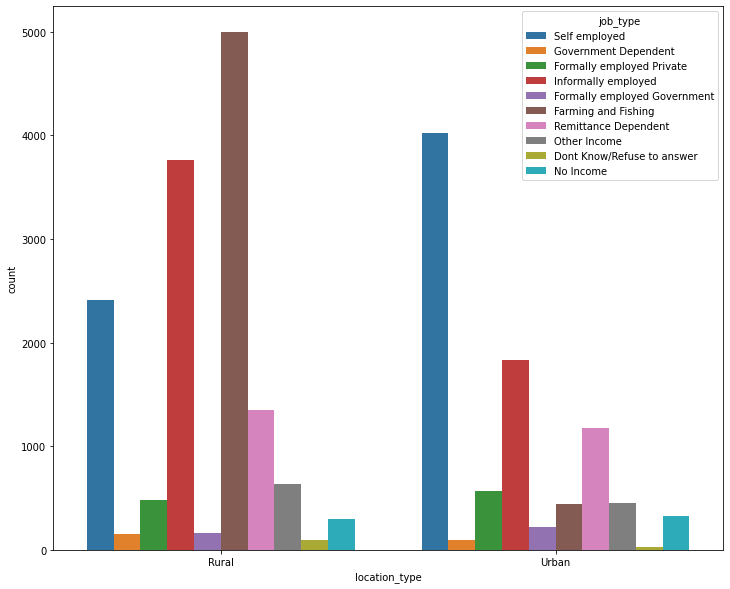
i)First the relation between cellphone access and job types count:

Here it was identified that the larger concentration of higher job types count was among those with cellphones, while there were substantially less job type counts for those without cellphone access.

.

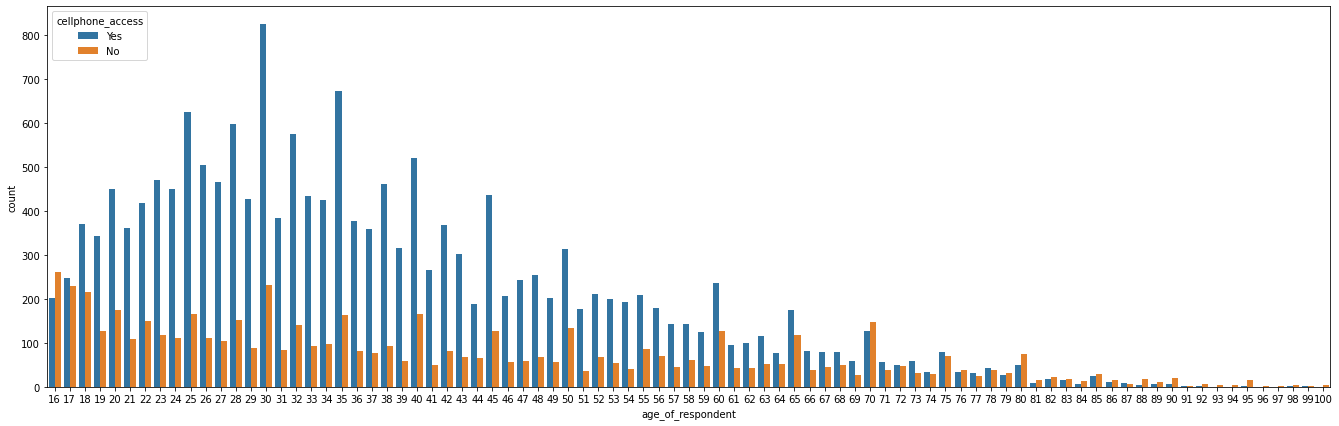
ii) Second the relation between respondent location type and job types count:

Here it was identified that the larger concentration of higher job types count was among those with cellphones, while there were substantially less job type counts for those without cellphone access.



**To determine relationship between cellphone access and respondent ages , then identification of the age has the most access to mobile phones**

The trend was identified as follows; a sharp increase between the 16 years to 30 years, the steady and slightly steep decline to the 80 years before reaching near zero.



The 30 year group has a number of 824 cellphone accessing persons, this is the largest number of cellphones for any group.

**Determine which age has the highest access to mobile phones and bank accounts.**

The 30 year age group has the highest number of persons with both bank accounts and cellphone access.This number being 176.

**To determine the percentage of level of access to mobile phones for people in the region per country.**

The percentage values for cellphones per region is done in relation to the total number of cellphones (17,454).

Country Cellphone Count Percentage

Rwanda 7249 41%

Kenya 4785 27%

Tanzania 3956 22%

Uganda 1464 8%

**To determine the percentage of level of access to mobile phones for people in the region.**

From a total of 23,525 persons, 17,454 people have been identified as having cellphones. This bring the cellphone percentage coverage to 74%.

**To determine the percentage of level of access to mobile phones for people in urban areas and rural areas**

From an investigation into distribution of cellphones (being a total of 17,454) in urban and rural regions. It has been noted that 11,070 people with phones are from rural areas ,while 6,384 of the people with cellphone are from the urban areas.This comes to a percentage coverage of 63% for rural areas and 36% for urban areas.

The above analysis was done using SQLite. The full analysis can be found in the following notebook.[[Link](https://colab.research.google.com/drive/13R2HDP4TGljgE0GdkXr5PU8ZWFlbN-1f?usp=sharing)]

## **5. Recommendations**

From our analysis, we would recommend that the fintech companies should target people in the age group of 18 to 40 as the market for their products.The banks should target the self employed people in farming and fishing industries for mobile banking integration.Also, the fintech companies should focus more on rural areas because there is a bigger market which has not been tapped.Therefore, prioritizing the recommendations above would not only increase the return on investment but ensure financial inclusivity across the East African region.

**Project Reference links**

Project management software used: [[Link](https://kevo.atlassian.net/jira/software/projects/BB/boards/2)]

Github: [[Link](https://kevo.atlassian.net/jira/software/projects/BB/boards/2)]

SQL analysis (colab notebook):[[Link](https://colab.research.google.com/drive/13R2HDP4TGljgE0GdkXr5PU8ZWFlbN-1f?usp=sharing)]

Python analysis and Vizualization(colab notebook) :[[Link](https://colab.research.google.com/drive/1lOinLHShfZT2ybsbQDvhnfOtG04nqcYo?usp=sharing)]