Business Understanding

Business Overview

Karamoja is the most food insecure region of Uganda. One of the main reasons is due to the low productivity level of the crops due to intense droughts as well as pest disease outbreaks. Several NGOs provide technical support as well as farm inputs to farmers experiencing low yield.Dalberg Data Insights (DDI) has been requested to develop a new food security monitoring tool to support the decision making of one of those NGOs active in Karamoja.To do so, Dalberg Data Insights developed a methodology to remotely measure the yield of the two main staple crops of the region (i.e. sorghum and maize) based on satellite images. The agri-tech team just ran the model for the 2017 crop season.

Business Objectives

To develop a visualization tool of the results of the first crop season. This visualization tool will be used as a model of the food security modeling tool that DDI will develop for the NGO.

Business Success Criteria

To be able to find patterns in the crop season, find trends and information on the districts and sub county level.

Assessing the situation

Requirements

1. **Resources**

- Data, we have the following zip folder link
- Personnel
- Software (Google Docs, Tableau)

2. **Assumptions**

- The data is correct and accurate
- The OBJECTID in both datasets are the same

3. **Constraints**

Tableau puts strain on the cpu hence the computer might hang and become slow while working.

4. **Cost / Benefit analysis**

The benefit of this project is that if it is done well and a good model is made out of our analysis, then we will get funding from the NGO however , we will need to pay all the data scientists working on this project with us. Further analysis can be done and a Cost/Benefit report can be given.

Data Mining goals

Our data mining goal for this project is to see which districts in karamoja are struggling the most, in terms of productivity and yield.

Potential questions include:

- Which districts are less productive?

- Which district has the least yield ?
- Which sub county has the least yield?
- What is the yield per acre?
- What is the yield in relation to population?

Project Plan

The CRISP-DM will be used as a guideline for this project

Data Understanding

Data understanding overview

The dataset contained in the zip files have sample data collected about sorghum and maize in the karamoja area in Uganda. The data files we will be using for this project are

1. Uganda Karamoja District crop yield population
2. Uganda Karamoja Subcounty crop yield population

Collecting Initial Data

The data was collected by the DDI.

Describing and exploring Data

We used tableau to explore the data and describe them. There are two datasets, both have information about the sorghum and maize plant however, one is for the district level and the other is for the sub county level.

Uganda Karamoja District crop yield population

Some of the column names include:

- POP: total population for the sub county
- S_Yield_Ha: average yield for sorghum for the district (Kg/Ha)
- M_Yield_Ha: average yield for maize for the district (Kg/Ha)
- Crop_Area_Ha: total crop area for the district (Ha)
- S_Area_Ha: total sorghum crop area for the district (Ha)
- M_Area_Ha: total maize crop area for the district(Ha)
- S_Prod_Tot: total productivity for the sorghum for the district (Kg)
- M_Prod_Tot: total productivity for the maize for the district (Kg)

Uganda Karamoja sub county crop yield population

POP: total population for the sub county

S_Yield_Ha: average yield for sorghum for the sub county (Kg/Ha)

M_Yield_Ha: average yield for maize for the sub county (Kg/Ha)

Crop_Area_Ha: total crop area for the sub county (Ha)

S_Area_Ha: total sorghum crop area for the sub county (Ha)

M_Area_Ha: total maize crop area for the sub county (Ha)

S_Prod_Tot: total productivity for the sorghum for the sub county (Kg)

M_Prod_Tot: total productivity for the maize for the sub county (Kg)

Verifying Data Quality

The data does not contain any missing values.However I had to add a latitude and longitude column for mapping issues.

Data Preparation

Selecting Data

From the zip folder, we selected:

Uganda Karamoja District crop yield population

Uganda Karamoja Subcounty crop yield population **,** as they were the most relevant in our analysis, used tableau to open the data and viewed it there.

Data Cleaning

Data cleaning procedures performed during analysis include:

- Renamed the columns to the full names.

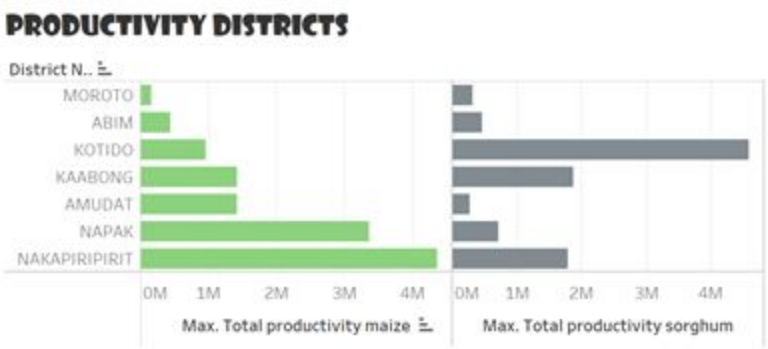
Integrating and formatting data

New data was constructed upon joining the two datasets on the OBJECTID column.

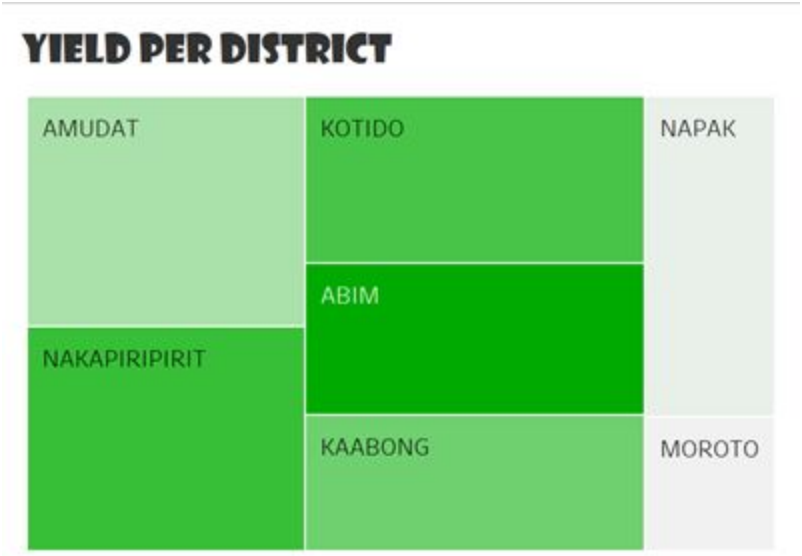
Analysis

During the analysis the following questions were answered through visualization

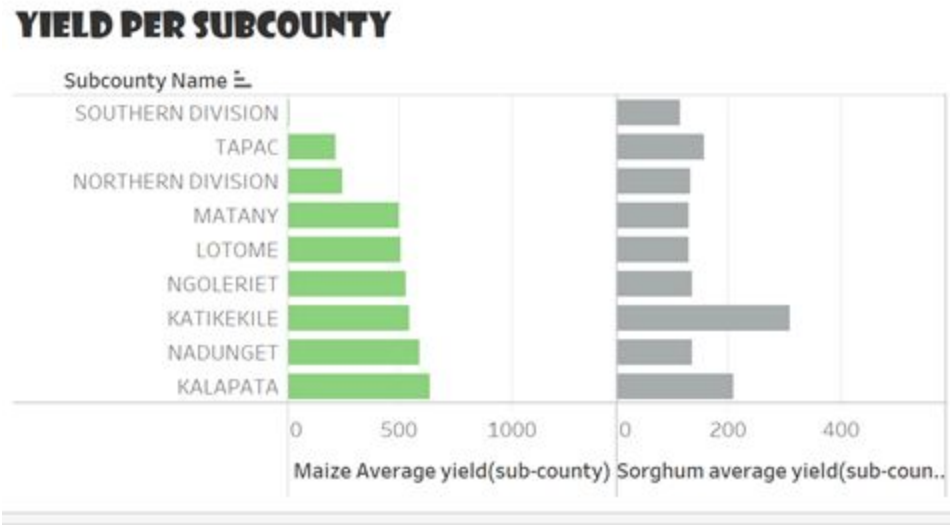
Which districts are less productive?



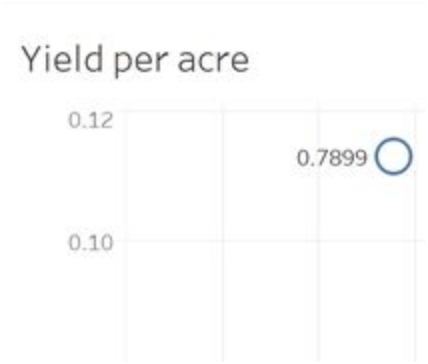
Which district has the least yield ?



Which sub county has the least yield?



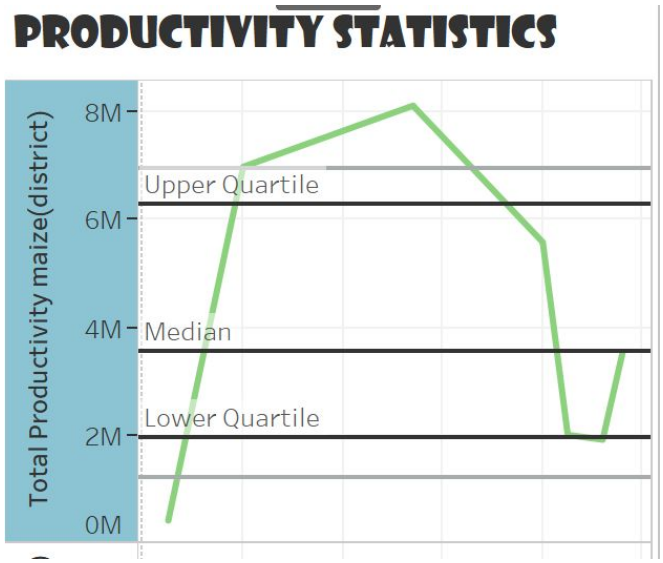
- What is the yield per acre?



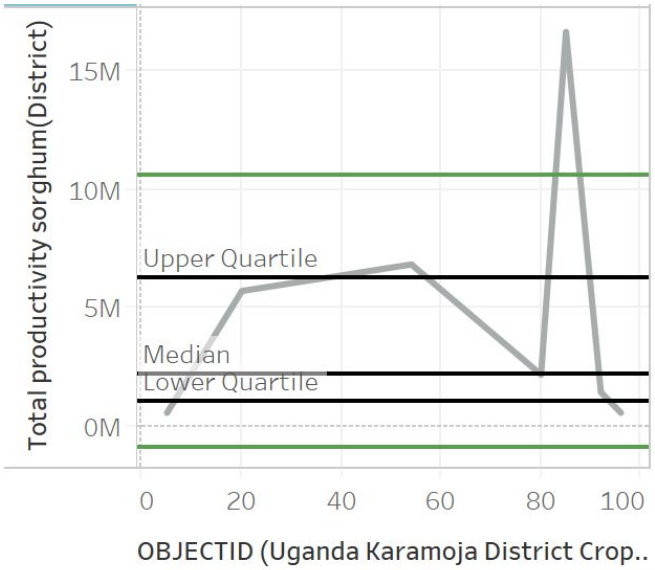
- What is the yield in relation to population?

We further analysed the data and upon further analysis we looked at the following questions:

Is the data skewed?

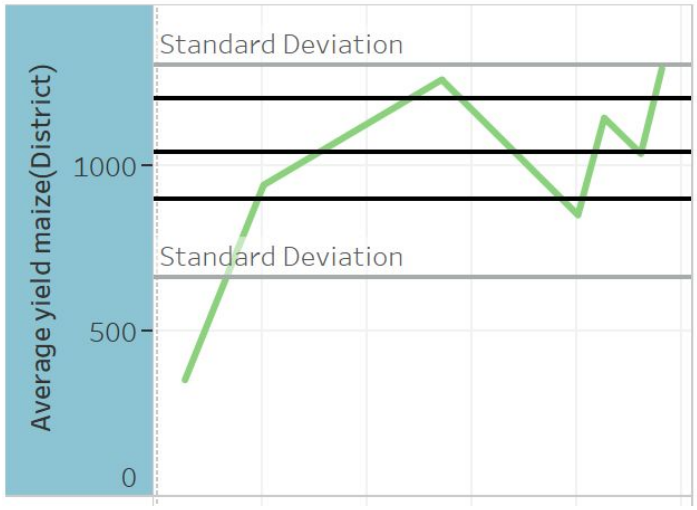


We can see that this is a positive skew. We can also see that there are a number of outliers in the dataset that lie outside the standard deviation.

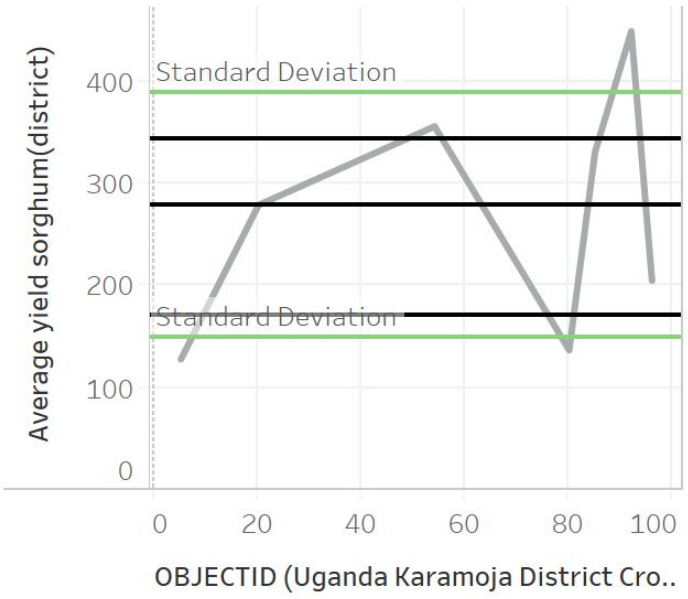


We can see that the skew is very positive. We can also see that there are outliers in the data that are outside the standard deviation.

YIELD STATISTICS

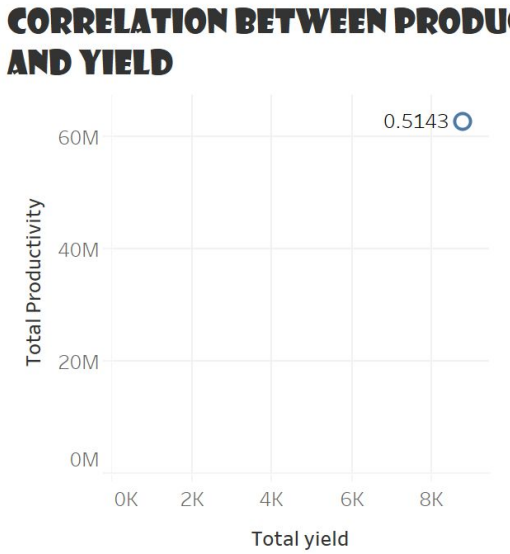


We can see that it is a normal distribution.However there are outliers in the data that are outside the standard deviation



We can see that there is a negative skew, however there are three outliers in the data that lie outside the standard deviation.

Is there a correlation between total productivity and total yield ?



To better see the visualisations you can visit the following page [Tableau]

Recommendation

My recommendation would be to begin with the Moroto and Abim district in the analysis considering the yield and productivity are positively correlated this means that they complement one another. Therefore Moroto and Abim are both unproductive and produce low yield hence they are in most need of the food.

Evaluation

According to the business success criteria, we were able to identify the most important information on the district and sub county level which is the yield and productivity for the maize and sorghum.