# • A written description of your process and approach, including why you took the approaches you used

I started analyzing maize because it had 45569 observations which was the highest for all the crops. Also, maize is the only crop with reported survey data for all the countries (cimbing bean just Burundi and Rwanda).

Initially, I explored possible predictors variables plotting them against yield. I converted all the yield to kg/ha to have more observations. I did something similar for the fertilizers.

# • Summary of any findings, even if preliminary

Chart, box and whisker chart

Description automatically generated

**Figure 1.** Yield (kg/ha) for each crop and country.

When the farmers report drought the yield tends to be lower. However, it does not look to be an effect if the farmers have electricity or not on yields (figure 2).

Chart, box and whisker chart

Description automatically generated

**Figure 2.** Yield (kg/ha) of maize when there is drought (1 in x axis) for farmers that report not having electricity (first block of data with 0 in the header) or having it (1 in the header).

I evaluated the relationship of fertilizer application with maize yield. However, as reported in the description of the dataset, there are some values that could not be reliable, e.g., more than 500 kg/ha of NPK reported in Zambia. Also, except for Kenya, Zambia and Tanzania most of farmers do not apply fertilizers. Zambia shows some increase in yield when more urea is applied to the field.

Chart, scatter chart

Description automatically generated

**Figure 3.** Yield (kg/ha) of maize based on NPK fertilizer applied to the field by country.

Chart, scatter chart

Description automatically generated

**Figure 4.** Yield (kg/ha) of maize based on urea fertilizer applied to the field by country.

The application or not of compost also does not look to affect the yield of maize.

Chart, box and whisker chart

Description automatically generated

**Figure 5.** Yield (kg/ha) of maize for farmers that reported using compost (1) or not (0).

Verifying if weeding influences yield, there is not a clear pattern identified (Figure 4).

Chart, histogram

Description automatically generated

**Figure 6.** Yield (kg/ha) of maize based on the number of times the farmer reported weeding the field.

Farmers that have cows then to have higher yield of maize.

Chart, box and whisker chart

Description automatically generated

**Figure 7.** Yield (kg/ha) of maize when farmer reported having cows (1) or not (0).

In contrast, farmers that have goats show a slightly reduction on yield. Do the goats eat maize? They do for cassava leaves, affecting the yield.

Chart, box and whisker chart

Description automatically generated

**Figure 8.** Yield (kg/ha) of maize when farmer reported having goats (1) or not (0).

The use of hybrid seeds contributes to higher yield.

Chart, box and whisker chart

Description automatically generated

**Figure 9.** Yield (kg/ha) of maize when farmer reported using hybrid seeds (1) or not (0).

Farmers that intercrop tend to have slightly lower yields.

Chart, box and whisker chart

Description automatically generated

**Figure 10.** Yield (kg/ha) of maize when farmer reported intercropping (1) or not (0).

The number of household members does not look to be related to yield.

Chart, scatter chart

Description automatically generated

**Figure 11.** Yield (kg/ha) of maize vs. number of household members for each country.

There is not clear effect of the effect of pests in the yield of maize.

Chart, box and whisker chart

Description automatically generated

**Figure 12.** Yield (kg/ha) of maize when pest and diseases were reported (1) or not (0).

There is not clear effect of the planting distance on the yield.

Chart, scatter chart

Description automatically generated

**Figure 13.** Yield (kg/ha) of maize vs. planting distance.

Adjusting a linear model with the variables selected (intercrop,hybrid,cows\_binary, urea\_kg\_ph\_all, npk\_kg\_ph\_all, drought, electricity, yield\_kg\_ph\_all, country,year, yield\_kg\_ph\_all) we are just able to explain 7.8% of the variability of yield in maize.

Text

Description automatically generated

In addition, when removing all the NA values, we end it up with 5348 observations from the original of 45569 observations for maize. As result, the analysis just included two countries (Kenya and Rwanda). The result of the linear model shows than using hybrid seeds the yield can increase around 2000 kg/ha while farmers than experienced drought can suffer of a reduction in yield of around 1220 kg/ha. Intercropping can also reduce the yield by 435 kg/ha. Rwanda reported higher yields than Kenya.

Chart, scatter chart

Description automatically generated

**Figure 14.** Observed vs. simulated yield (kg/ha) of maize using a linear regression model.

I tried to include other variables as number of cows and the amount of hybrid seed used as predictor, but this reduced the number of observations without NAs to less than 2000.

The results of random forest for the variables selected does not improve the prediction of yield with a R2 of 0.0876.

Text

Description automatically generated

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

**Figure 15.** Observed vs. simulated yield (kg/ha) of maize using random forest.

# • Recommendations on any potential next steps for analysis

* Climbing beans could be the next crop to analyze with more surveys available after maize.
* Future analysis could consider split the information by country and explore other crops. For example, Rwanda that has beans, climbing beans, maize and potato which could allow a comparison among crops.
* The use of hybrid seeds needs to be corrected for some countries that did not ask this question for the 1AF farmers although they knew they used hybrid seed. However, probably was not reported in the final data sheet.
* Further analysis should include other predictors and review the missing data to see if it is possible to make some assumptions or avoid the number of observations removed due to the lack of reported answers for some of the questions in the survey.