Smart Farming

Predicting Crop Yield: Enhancing Agricultural Efficiency and Reducing Food Insecurity

Chidubem Nwabunze, Madathil Geetanjali Menon, Sai Madikonda, Aishwarya Sadagopan, Jasmine Hill

Table of contents

Overview

Overview – Our Goal

- Predict future crop yields of ten most consumed crops in the world to help farmers make informed decisions about planting and harvesting, and optimize crop yield
- Identify factors that impact crop yield and to what extent
- Find the model that accurately predicts crop yield depending on a set of influencing factors, such as country, rainfall, amount of pesticide use, etc.

Our Data

About Our Data

Data Source

Our dataset is sourced from Kaggle, which contains data from the FAO and the World Data Bank. We collected additional data from the FAO to add to the dataset

Sample Size (*n*)

Our cleaned dataset has 26,092 data points

Number of Variables (*k*)

Our dataset contains 11 variables

Our data

Why We Chose This Data

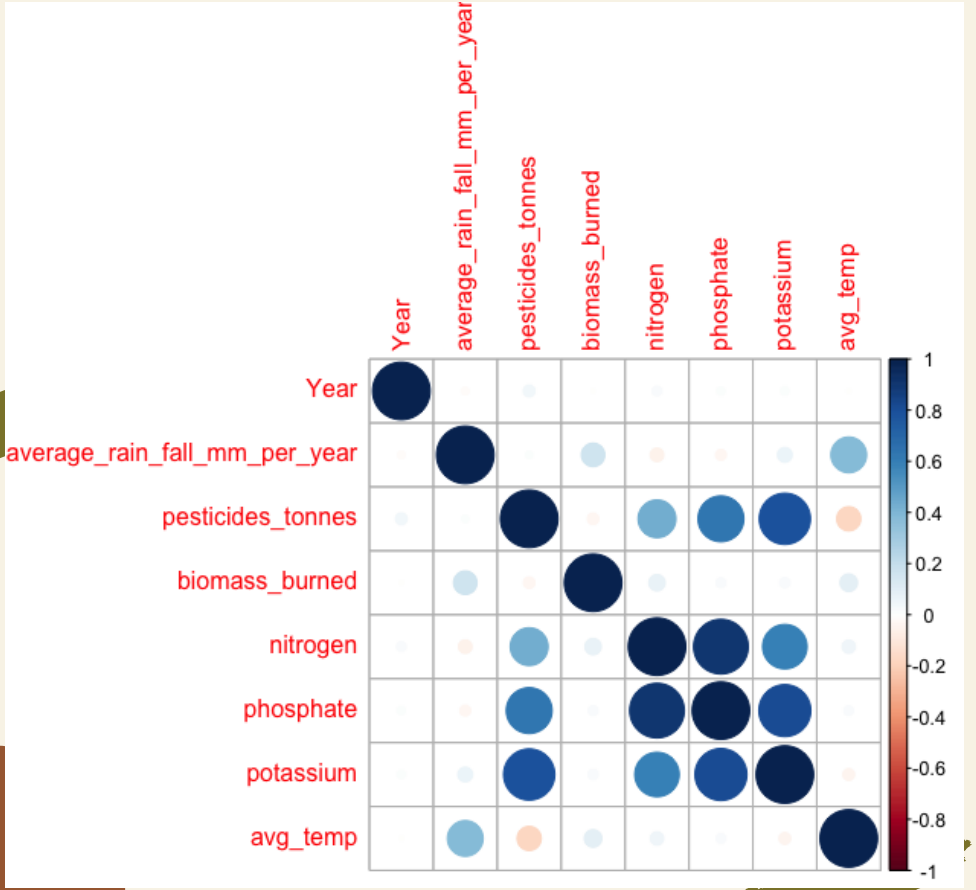
This dataset contains many variables that may influence crop yields of the ten most consumed crops in the world. Our predictions will be applicable to a larger population of farmers worldwide because these crops are produced in great quantities to fulfill demand across borders.

About Our Data – Details

Variable	Description	Type
hg/ha Yield	The dependent variable and is measured in hectograms per hectare (hg/ha)	Numerical
Area	The country where the crop is produced from	Categorical
Item	The crop produced	Categorical
Year	The year when the crop was grown	Categorical
Average Rainfall	The average rainfall measured in millimeters per year (mm/year)	Numerical

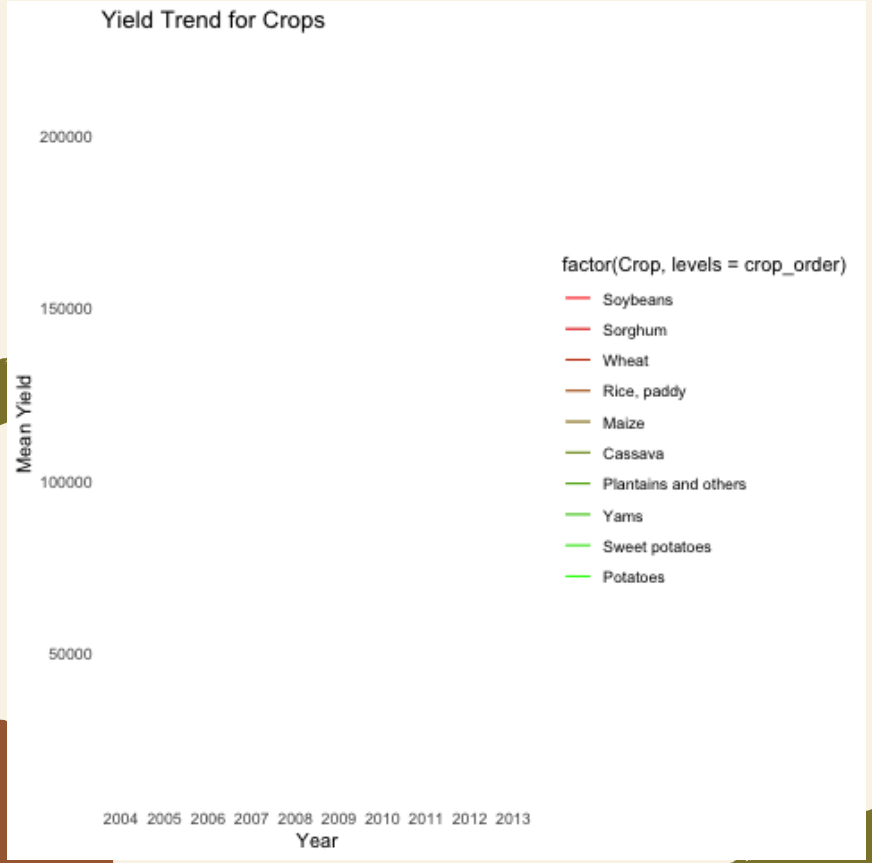
Crop	Description	Type
Pesticides	The amount of pesticides used measured in tonnes	Numerical
Average Temperature	The average temperature when the crop was grown	Numerical
Nitrogen	The amount of nitrogen in the soil measured in tonnes	Numerical
Phosphate	The amount of phosphate in the soil measured in tonnes	Numerical
Potassium	The amount of potassium in the soil measured in tonnes	Numerical
Biomass Burned	The amount of biomass burned in organic soils from wildfires reported in tonnes	Numerical

Correlation Plot



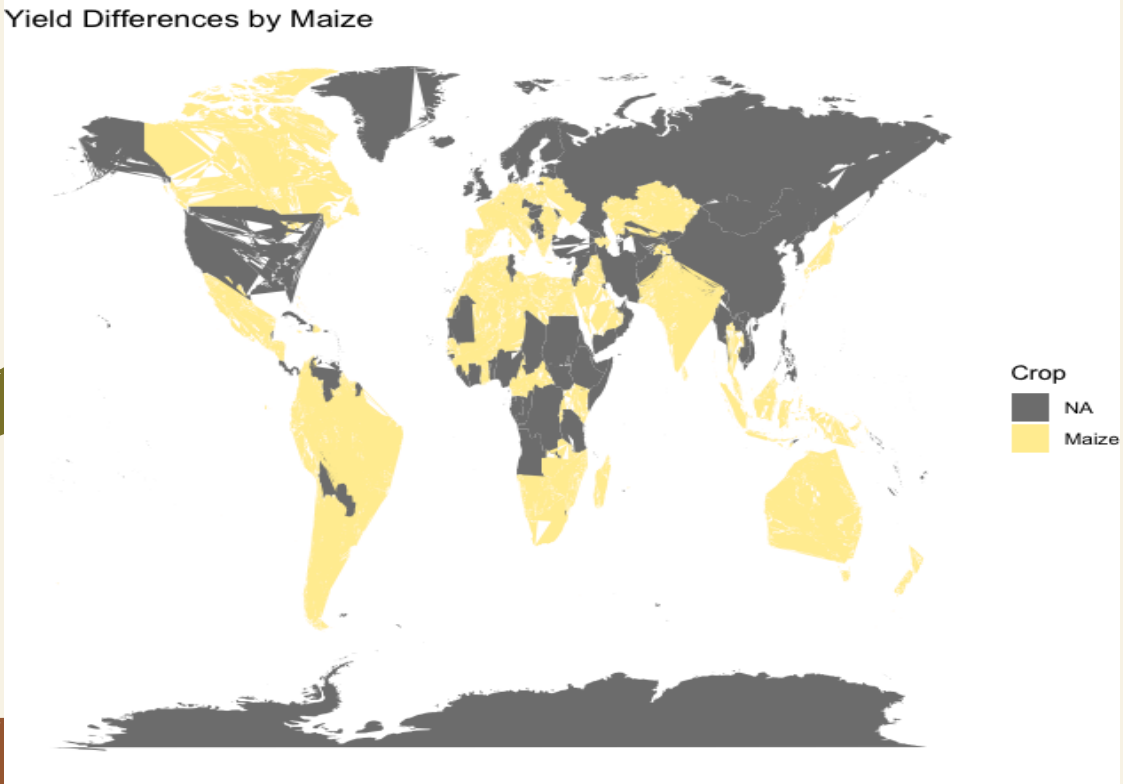
A graphical representation of the correlation between variables, helping to identify the strength and direction of their relationship.

Yield Trend for crops



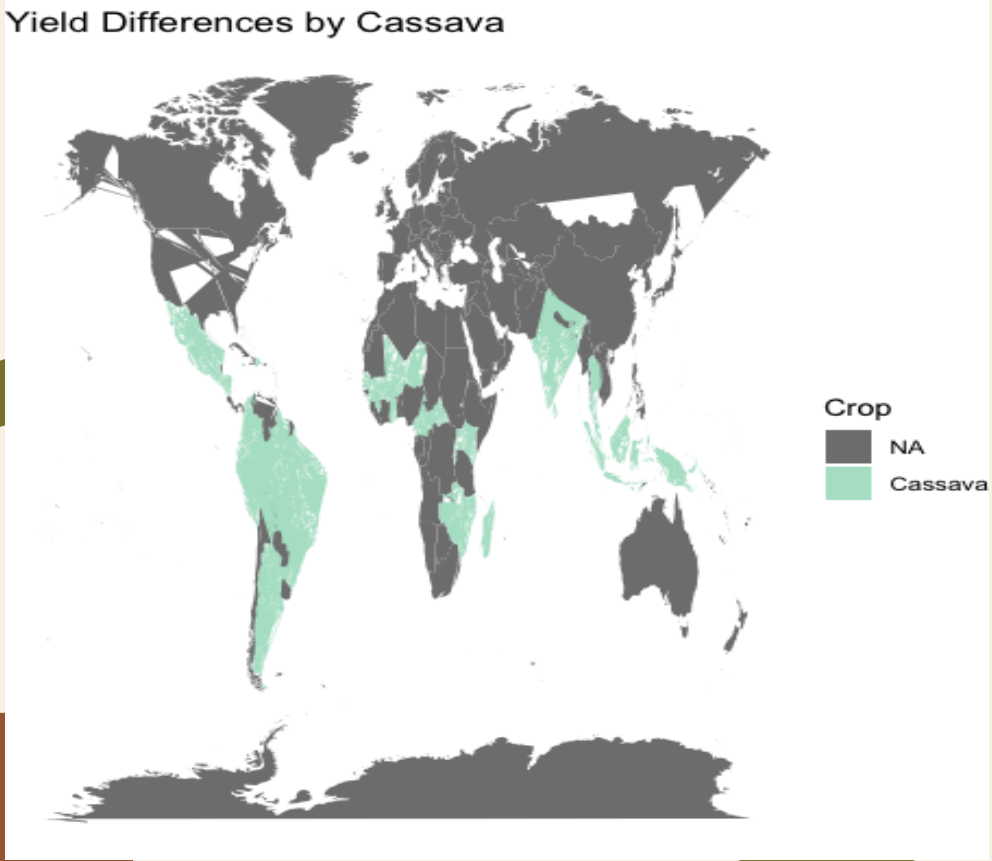
An analysis of the production trend of a particular crop over a period of time, providing insights into the crop's performance and potential factors affecting its yield.

Yield Differences by Maize



A visual representation of the variations in maize production across different regions and countries, highlighting potential factors affecting yield differences

Yield Differences by Cassava



A visual representation of the variations in cassava production across different regions and countries, highlighting potential factors affecting yield differences.

Methodology

Methodology

Analysis & Results

Linear Regression – Analysis and Results

- Employed for predicting area yield for different crops due to its suitability for modelling the linear relationship between independent variables
- Partitioned data before and after 2004 into training and testing data respectively
- Plotted a correlation matrix to remove multicollinearity

RMSE Values

Training Set	Test Set
37935.87	46238.99

Regression Tree – Analysis and Results

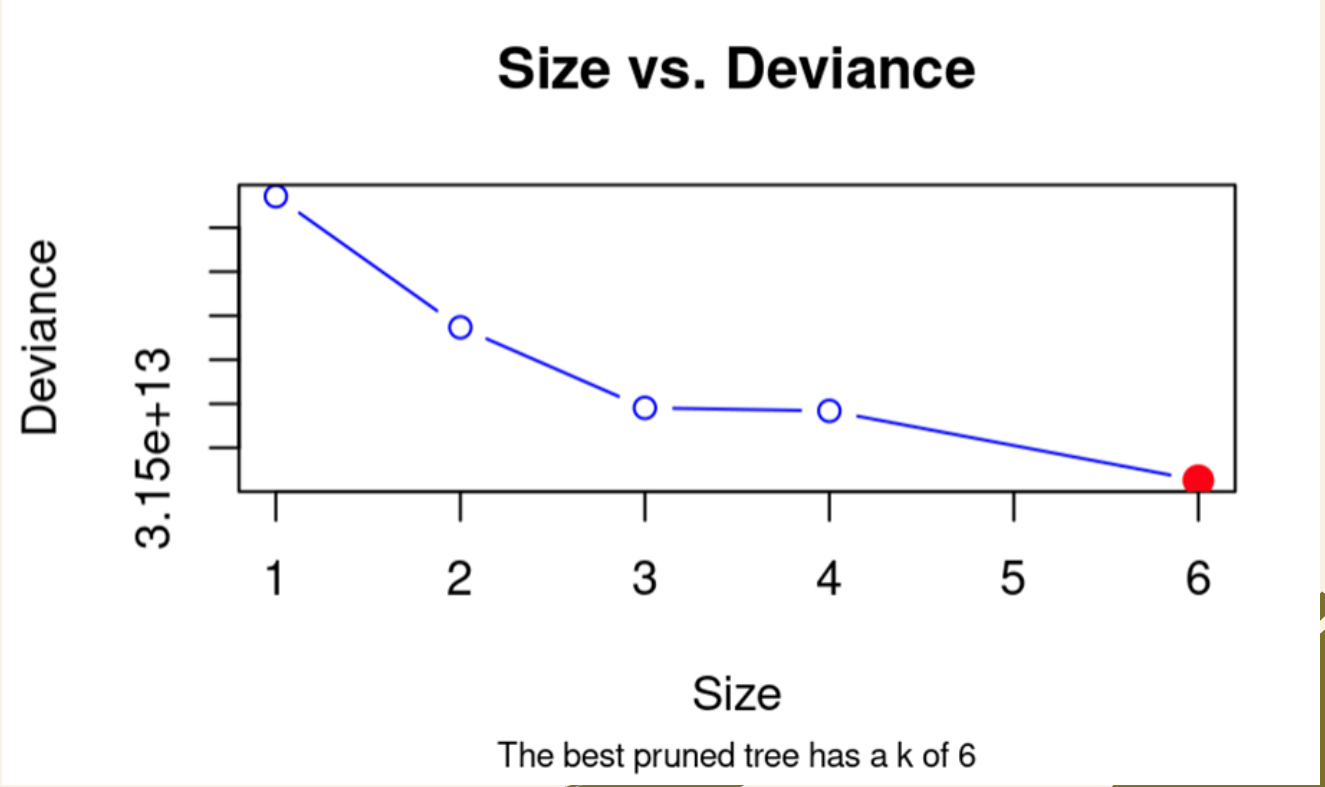
- Used to predict the yield of crops based on several numerical variables
- Resulted in six terminal nodes

RMSE Values

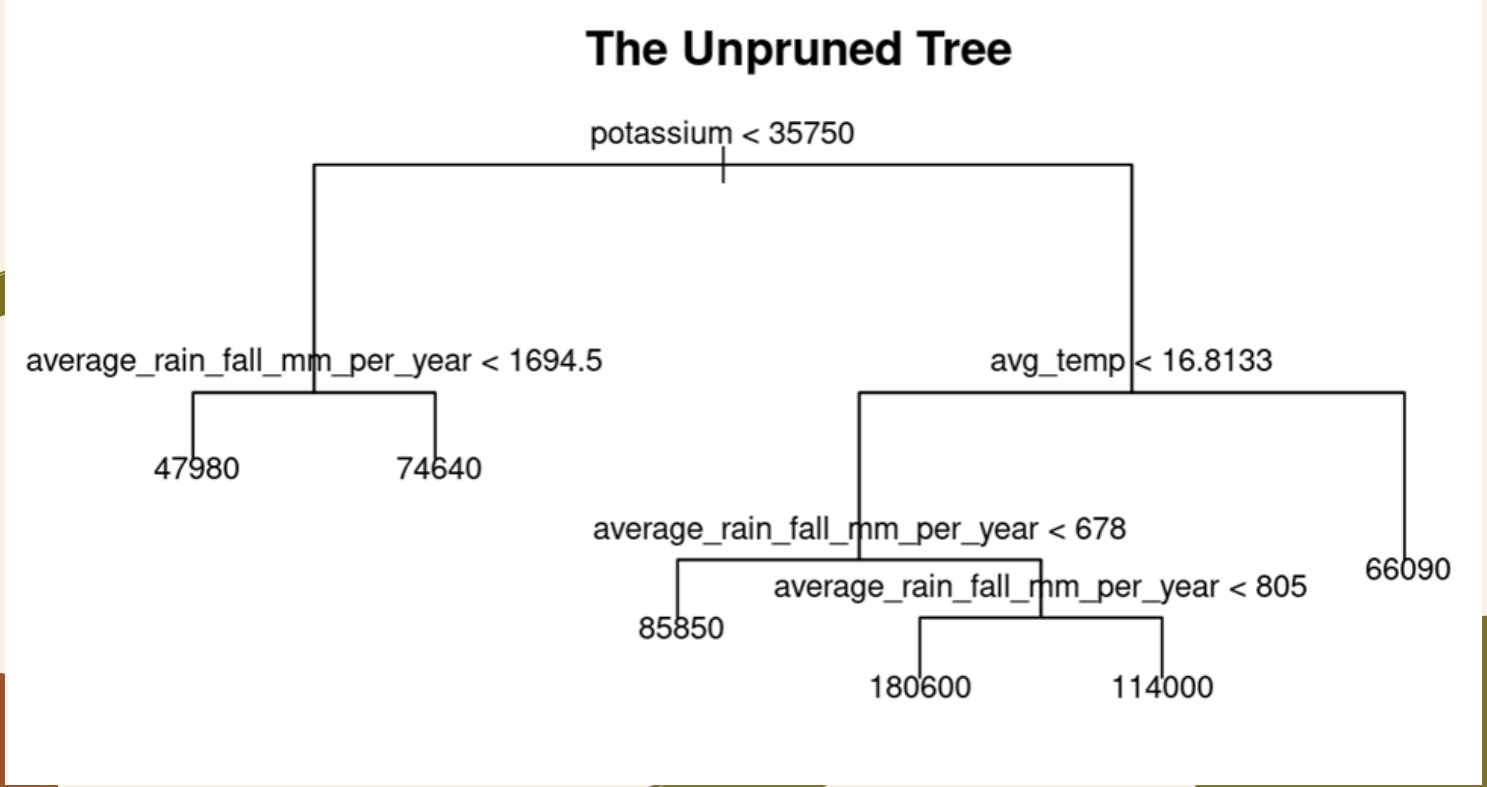
Training Set	Test Set	Pruned Test Set
69,937.74	83,607.50	83,607.50

Minimum Yield	Maximum Yield
50 hg/ha	501,412 hg/ha

Regression Tree – Analysis and Results



Regression Tree – Analysis and Results



XG Boost Method – Analysis and Results

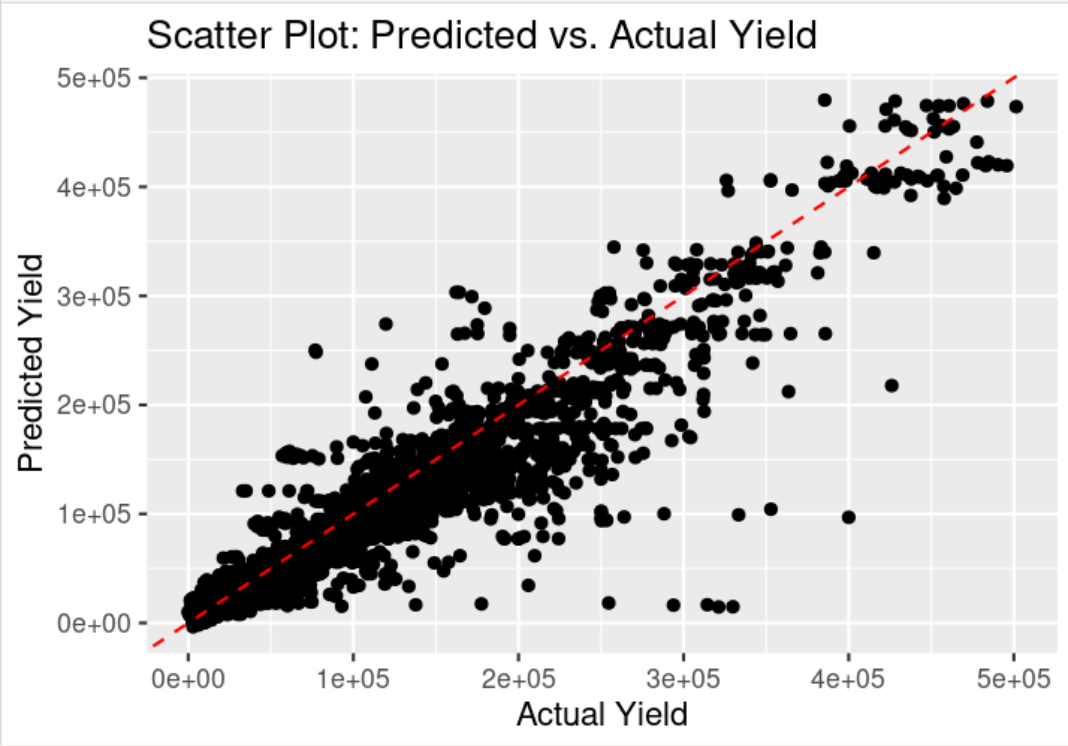
- A XGBoost model was used to predict crop yield using various variables such as country, crop, year, rainfall, pesticides, biomass burned, and nutrients
- The data was split into training and testing sets based on the year, and unique countries in the testing set were identified and those countries not present in the training data were removed
- The RMSE on the training data was 32832.28, while the RMSE on the testing data was 47008.54.

Gradient Boost Method – Analysis and Results

- The GBM model was trained using the training data, with a specified distribution of "gaussian" and a total of 5000 boosting trees. The interaction depth was set to 4 and shrinkage was set to 0.1

- The RMSE on the training data was determined to be 10759.85 while the RMSE on test data was found out to be 27325.66

Gradient Boosting Method – Analysis and Results



Clustering – Analysis and Results

- We considered two clustering algorithms: K-means Clustering and Hierarchical Clustering.
- Using K-means Clustering Algorithm, our goal was to predict average crop yield of different clusters and to identify the cluster with the highest average crop yield as well as the cluster members.
- Upon close inspection we noticed that the difference in crop yield and other features were minimal across the different years for the different countries. We then decided to group the data by crop and country, by taking an average of all predictors for the different years and dropping the year column.

  There were ten different crops in our dataset for the clustering analysis we decided to analyse **Maize**, **Potato**, **Wheat**, and **Cassava** only.
- For each of the crop, using the elbow method, K = 4 seemed to be the optimal number of clusters.
- We were able to use K-means Clustering Algorithm to predict average crop yield of different clusters and to identify the cluster with the highest average crop yield as well as the cluster members.

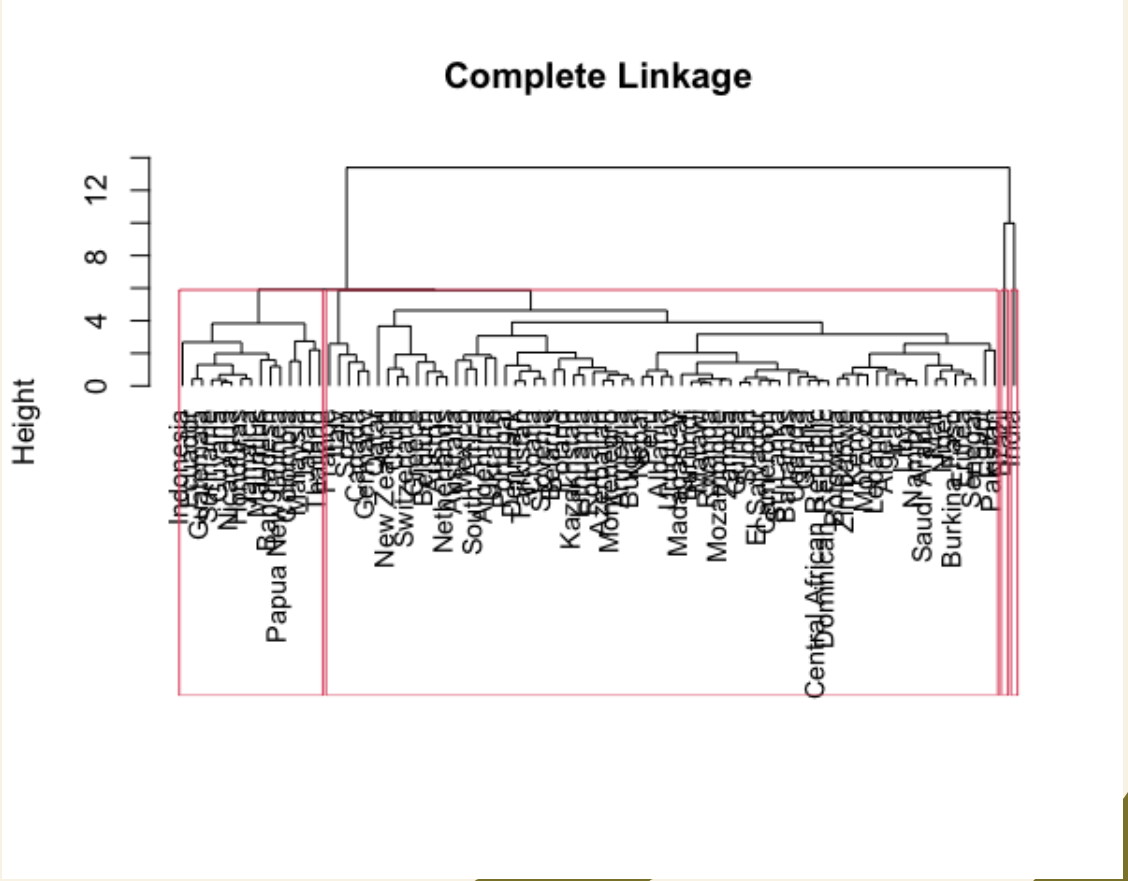
Clustering – Analysis and Results

K-Means Clustering Results of Cluster with highest average yield			
Crop	Cluster No.	Cluster Members	Average Yield
Maize	Cluster 3	Argentina, Armenia, Australia, Austria, Azerbaijan, Belarus, Belgium, Bulgaria, Canada, Chile, Croatia, Denmark, France, Germany, Italy, Japan, Kazakhstan, Lithuania, Mexico, Montenegro, Netherlands, New Zealand, Poland, Romania, Slovenia, Spain, Switzerland, Tajikistan, Ukraine.	21308.21 hg/ha
Potato	Cluster 3	Argentina, Armenia, Australia, Austria, Azerbaijan, Belarus, Belgium, Bulgaria, Canada, Chile, Croatia, Denmark, Estonia, Finland, France, Germany, Ireland, Italy, Japan, Kazakhstan, Latvia, Lithuania, Mexico, Montenegro, Netherlands, Norway, Poland, Portugal, Romania, Slovenia, Spain, Sweden, Switzerland, Tajikistan, Ukraine.	93151.62 hg/ha

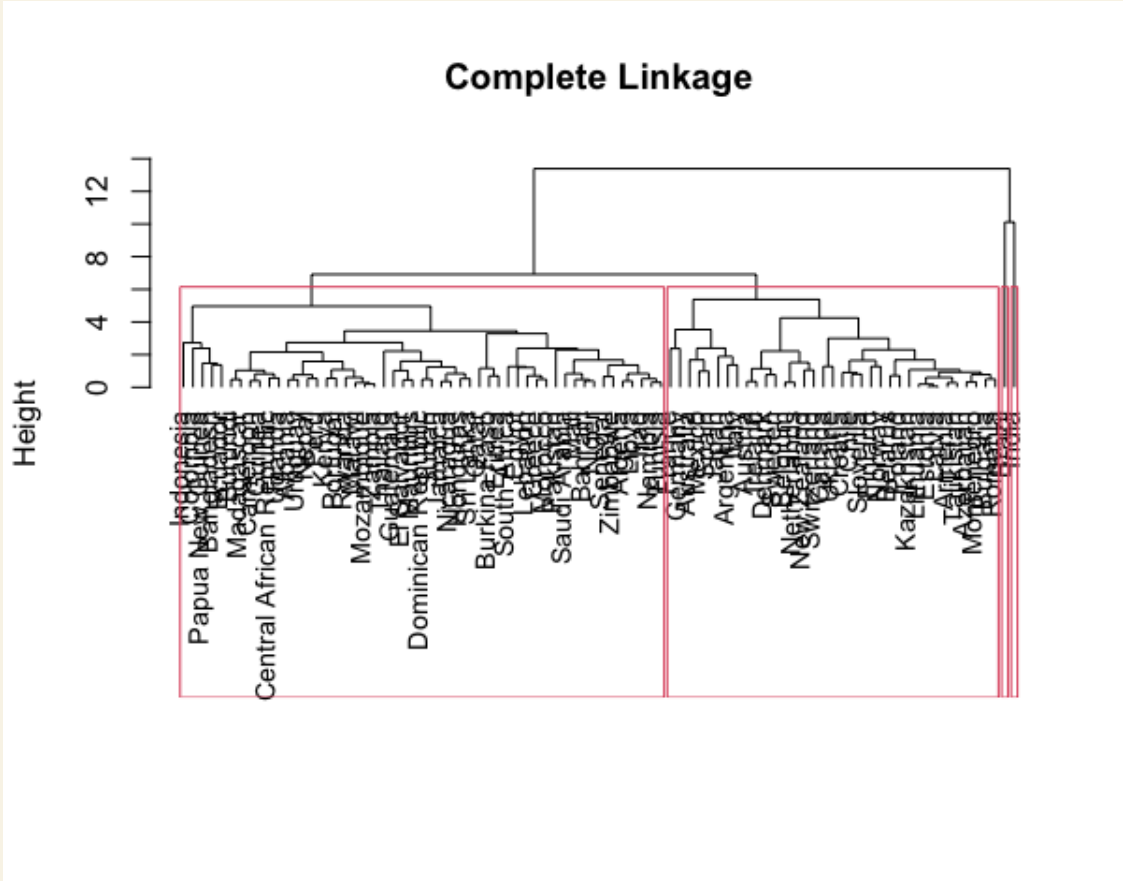
Clustering – Analysis and Results

K-Means Clustering Results of Cluster with highest average yield			
Crop	Cluster No.	Cluster Members	Average Yield
Wheat	Cluster 3	Armenia, Austria, Azerbaijan, Belarus, Belgium, Bulgaria, Canada, Chile, Croatia, Denmark, Estonia, Finland, France, Germany, Ireland, Italy, Kazakhstan, Latvia, Lithuania, Montenegro, Netherlands, Norway, Poland, Romania, Slovenia, Spain, Sweden, Switzerland, Tajikistan, Ukraine.	16848.52 hg/ha
Cassava	Cluster 1	Bahamas, Burkina Faso, Cameroon, Central African Republic, Dominican Republic, El Salvador, Ghana, Guinea, Mali, Niger, Senegal, Sri Lanka, Thailand, Uganda.	35816.37 hg/ha

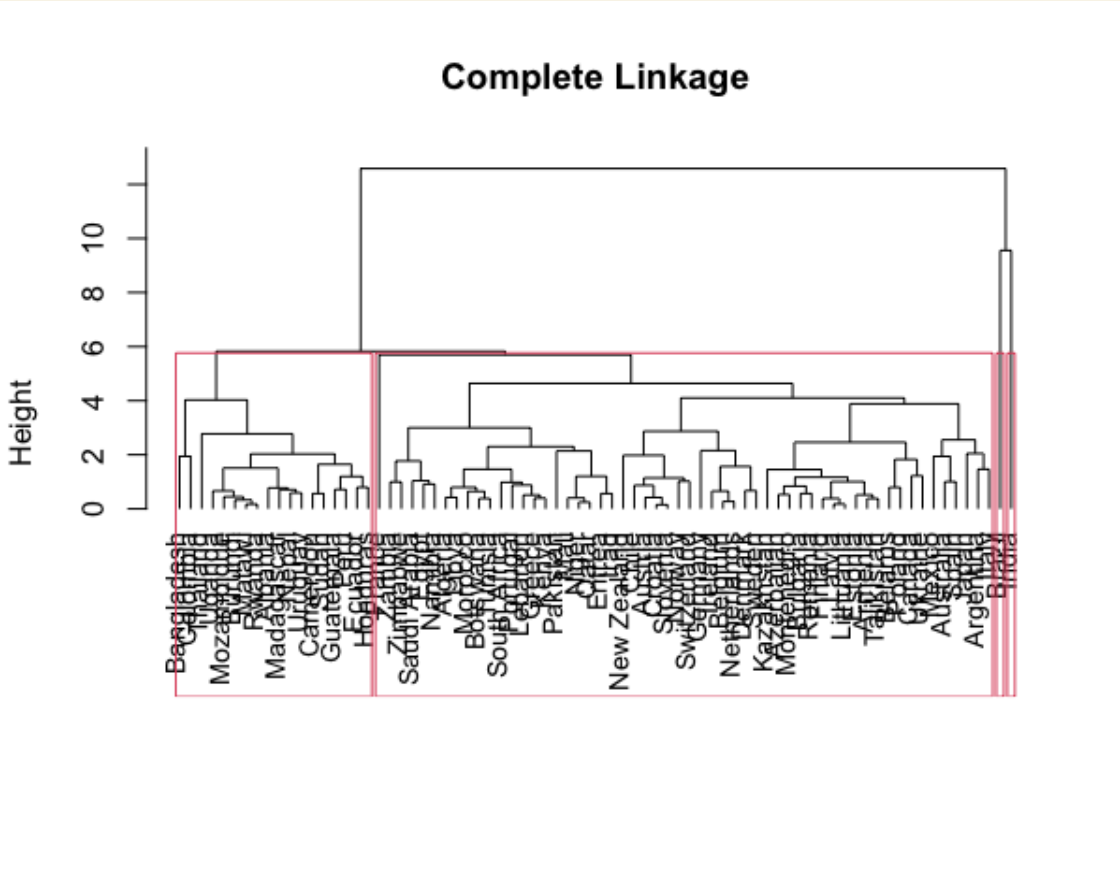
Hierarchical Clustering – Maize



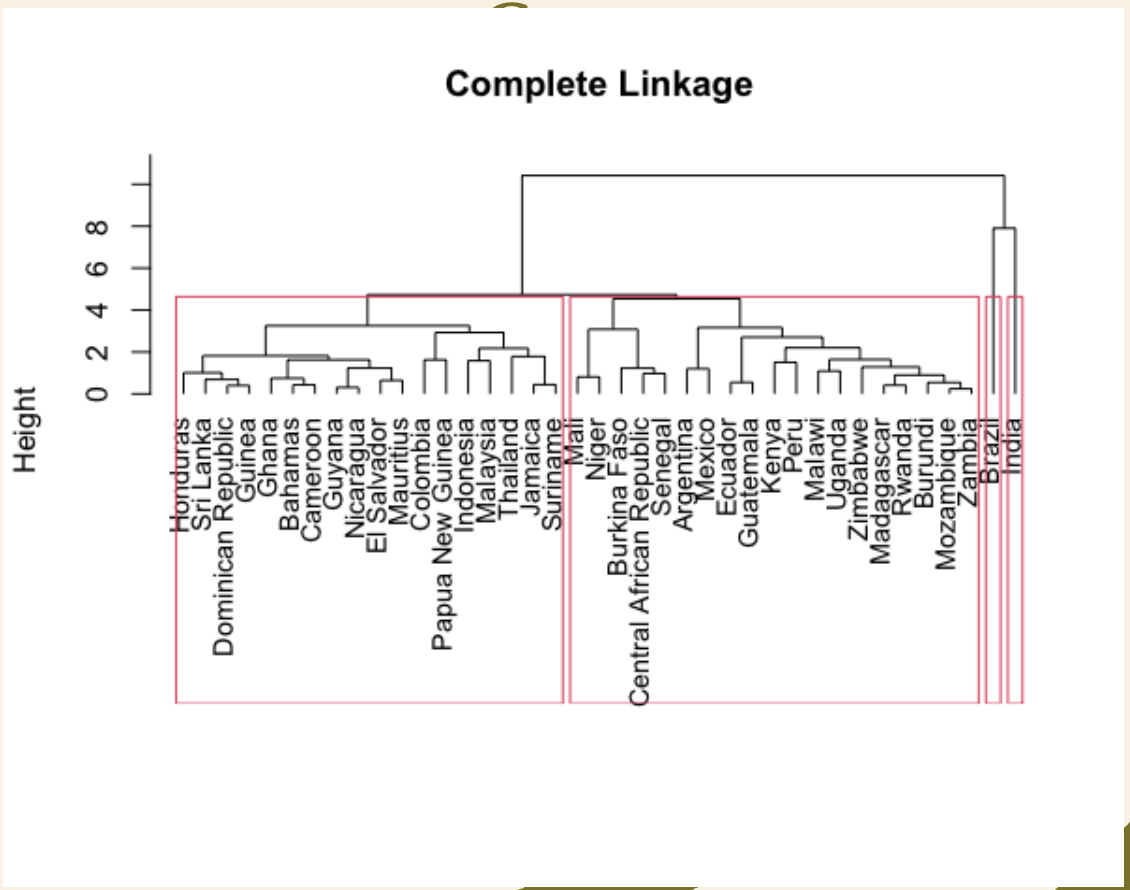
Hierarchical Clustering – Potato



Hierarchical Clustering – Wheat



Hierarchical Clustering –



Comparing Model Performance

- Gradient boosting is the model with the smallest test RMSE, indicating it has better predictive performance in comparison to regression trees and boosting methods

Linear Regression	Regression Tree	XGBM	GBM
46,238.99	84,181.4009	47,008.54	27,325.66

Conclusion

Conclusion – Our Recommendations

- Predicting crop production is essential to a farmer's capacity to make a profit. In order to create a prediction model that can assist farmers in making wise decisions about planting and harvesting, in this this project we have used R to examine historical data on crop yields and weather trends.

- Four models were used to make predictions and their RMSE values were calculated: linear regression (RMSE=46238.99), regression tree (RMSE=84181.4009), XGBM (RMSE=47008.54), and GBM (RMSE=27325.66).

- The results suggest that GBM had the lowest RMSE, followed by XGBM, linear regression, and regression tree, indicating that **Gradient Boosting Method** was the most accurate model for making predictions.

Thanks!

Any Questions?