

FORECASTING FERTILIZER EFFICIENCY AND AGRICULTURAL PRODUCTIVITY DATA REPORT

MORINGA DSF-FT 11 – HYBRID

GROUP II PROJECT

PRESENTED BY:

Rodgers Ndemo

Petina Leni

Kennedy Kariuki

Betsy Gitje

Michael Gatero

Table of Contents

FORECASTING FERTILIZER EFFICIENCY AND AGRICULTURAL PRODUCTIVITY DATA REPORT

1.0 Business Understanding	1
1.1 Business Context	1
1.2 Business Problem	1
1.3 Project Objectives	1
1.3.1 Main Objectives	1
1.3.2 Specific Objectives	1
1.4 Success Metrics.....	2
1.5 Stakeholders	3
2.0: Data Understanding	4
2.1: Data Source	4
2.2 Features and Variables Description	4
2.3 Data Collection Methods	5
2.5 Data Quality Assessment	5
2.5.1 Missing Values	5
2.5.2 Outliers.....	5
2.5.3 Duplicates.....	5
2.5.4 Inconsistencies	6
3.0 Exploratory Data Analysis	7
3.1 Univariate Analysis	7
3.2 Bivariate Analysis.....	10
3.3 Multivariate Analysis.....	15
4.0 Modeling	18
4.1 Time Series Forecasting for Kenya’s Fertilizer Consumption.....	18
4.1.1 Data Preparation.....	18
4.1.2 Model Selection and Training	18
4.1.3 Forecasting Future Consumption (2024–2035)	18
4.1.4 Visualization	19
4.1.5 Interpretation of Forecast	19

4.2 Predicting Crop Production Index Using Regression Models	20
4.2.1 Feature and Target Selection	20
4.2.2 Handling Missing Values	20
4.2.3 Data Splitting and Standardization.....	20
4.2.4 Models Trained.....	20
4.2.5 Model Performance	21
4.2.6 Interpretation of Results	21
4.2.7 Final Model Selection	22
Recommendations	23
Conclusion	23

FORECASTING FERTILIZER EFFICIENCY AND AGRICULTURAL PRODUCTIVITY DATA REPORT

1.0 Business Understanding

1.1 Business Context

Agriculture forms the backbone of East Africa's economy, employing over 60% of the workforce and contributing significantly to the GDP. Despite investments in fertilizer programs, productivity remains low. This disparity raises questions about the efficiency and strategic application of fertilizers in the region.

1.2 Business Problem

Fertilizer use is rising across East Africa, but crop yield improvements are inconsistent and vary widely between countries. Policymakers, NGOs, and agritech investors are concerned that fertilizer is being applied inefficiently, without data-driven guidance. Without clear insights into the fertilizer productivity relationship and future needs, policies may misallocate resources, and farmers may suffer from suboptimal yields. There is need to optimize fertilizer use to sustainably boost agricultural output using historical data.

1.3 Project Objectives

1.3.1 Main Objectives

- To develop a data-driven framework that forecasts fertilizer usage and supports sustainable agricultural productivity across East Africa, empowering stakeholders with insights that guide better policies, investments, and resource allocation.

1.3.2 Specific Objectives

1. To analyze historical fertilizer consumption trends across East African countries from 1960 to 2023.
2. To investigate the relationship between fertilizer usage and agricultural productivity indicators.
3. To develop time-series models for forecasting future fertilizer demand up to 2035.

4. To build machine learning models that predict productivity outcomes based on fertilizer use and other variables.
5. To cluster countries with similar fertilizer efficiency patterns for targeted policy recommendations.
6. To generate insights and actionable strategies that support sustainable agriculture and food security beyond the scope of the project.

1.4 Success Metrics

1. Accuracy of Forecasting Models:

- Achieve a minimum accuracy of X% (e.g., >85%) in time-series forecasting models for fertilizer demand by 2035.
- Low Mean Absolute Percentage Error (MAPE) or Root Mean Square Error (RMSE) for predictions.

2. Correlation Strength:

- Strong positive correlation (R^2 value > Y, e.g., >0.7) between fertilizer usage and agricultural productivity indicators.

3. Cluster Performance:

- Clustering models should identify distinct groups of countries with > Z% homogeneity in fertilizer efficiency patterns.

4. Actionable Insights:

- Deliver at least 5 region-specific actionable policy recommendations supported by the analysis.
- Produce a comprehensive report used by at least 3 stakeholders (e.g., Ministries, NGOs, donors).

5. User Engagement:

- Stakeholder satisfaction score >80% based on post-project surveys.
- Adoption of the framework by at least 2 East African governments, donor agencies, or agritech companies.

6. Impact on Sustainable Development:

- Quantifiable increase in agricultural productivity or efficiency in pilot regions where recommendations are implemented (e.g., productivity increase by X metric tons/hectare).

7. **Visualization and Accessibility:**

- Interactive dashboards (Tableau or equivalent) with user-friendly designs evaluated positively by stakeholders.
- 100% of key findings accessible to stakeholders via reports, presentations, or dashboards.

1.5 Stakeholders

- **National Ministries of Agriculture** – for strategic input planning and subsidies.
- **Regional Bodies** (EAC, IGAD) – for coordination and policy harmonization.
- **Farmers' Cooperatives** – to optimize fertilizer application practices.
- **Agritech Companies** – to align product offerings with market needs.
- **Donors & NGOs** (FAO, World Bank) – for evaluating the impact of their interventions.
- **Investors** – identifying high-potential regions for agricultural investment.

2.0: Data Understanding

2.1: Data Source

The primary data source for this project is:

- **Humanitarian Data Exchange (HDX)** - [World Bank - Agriculture and Rural Development Dataset](#)

This dataset aggregates annual agricultural indicators from the **World Bank** and focuses specifically on rural development, agriculture inputs (like fertilizers), and output metrics (like crop yields).

The dataset is highly credible, updated regularly, and harmonized across countries and time periods, ensuring high levels of reliability and comparability.

Key advantages of the data source:

- Consistent global standards (World Bank methodologies)
- Wide temporal coverage (1960s–2023)
- Focused on agriculture and rural development metrics critical to our study

2.2 Features and Variables Description

Feature	Description
Year	The calendar year when the measurement was recorded. Annual frequency.
Country	The East African country to which the data record pertains.
Fertilizer Consumption (kg/ha)	Amount of fertilizer used per hectare of arable land. Reflects agricultural input intensity.
Cereal Yield (kg/ha)	Yield of cereal crops (e.g., maize, wheat, rice) measured in kilograms per hectare.
Arable Land (% of total)	Proportion of land classified as arable compared to the country's total land area.
Agricultural Land (% of land area)	Percentage of total land area dedicated to agricultural activities, including permanent crops.

Each of these variables was selected to reflect either **input factors** (fertilizer, land) or **output factors** (crop yield), forming the basis for predictive modeling and trend analysis.

2.3 Data Collection Methods

Data collection follows standardized protocols through:

- Annual surveys by national governments
- Remote sensing and satellite imagery (especially for land area measurements)
- Validation by World Bank experts to ensure harmonized reporting
- Historical interpolation for missing periods in some developing nations

The data undergoes a two-stage validation: first by national agencies, then by World Bank auditors before public release via HDX.

2.5 Data Quality Assessment

2.5.1 Missing Values

- Some countries have missing fertilizer consumption data, particularly in the 1960s-1980s.
- Cereal yield and land usage figures are more complete but occasionally missing during periods of political instability.
- Missing data handling strategies under consideration:
 - Linear interpolation for within-country gaps
 - Regional median imputation where interpolation is infeasible

2.5.2 Outliers

- Sharp increases in fertilizer use during specific years hint at external interventions (e.g., subsidy programs, donor support).
- Cereal yield outliers often correlate with major events like droughts, wars, or major agricultural reforms (e.g., Green Revolution uptake).

2.5.3 Duplicates

- No duplicate entries were found after initial inspection.

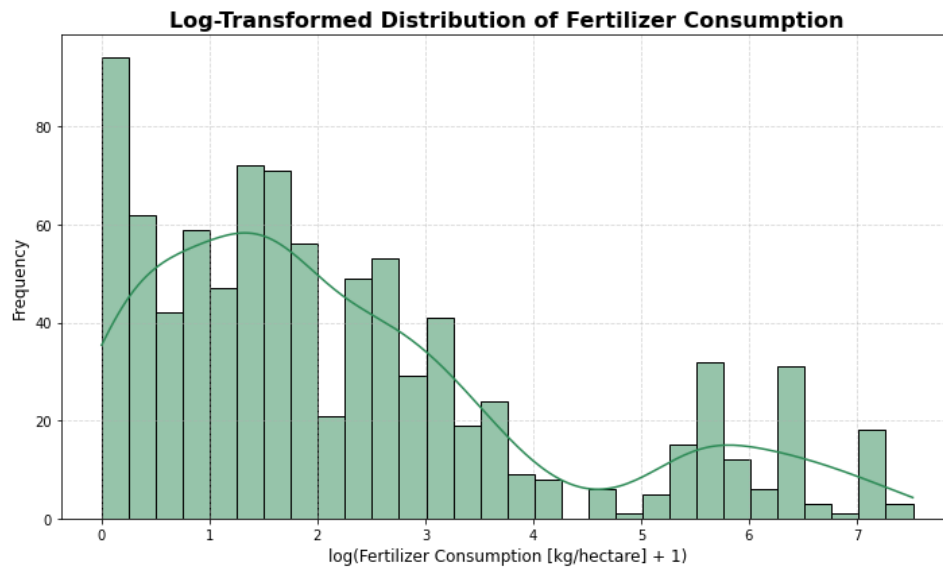
2.5.4 Inconsistencies

- Minor inconsistencies in country names (e.g., “Tanzania, United Republic of” normalized to “Tanzania”) were cleaned.
- Measurement units remained consistent across all countries and years.

3.0 Exploratory Data Analysis

3.1 Univariate Analysis

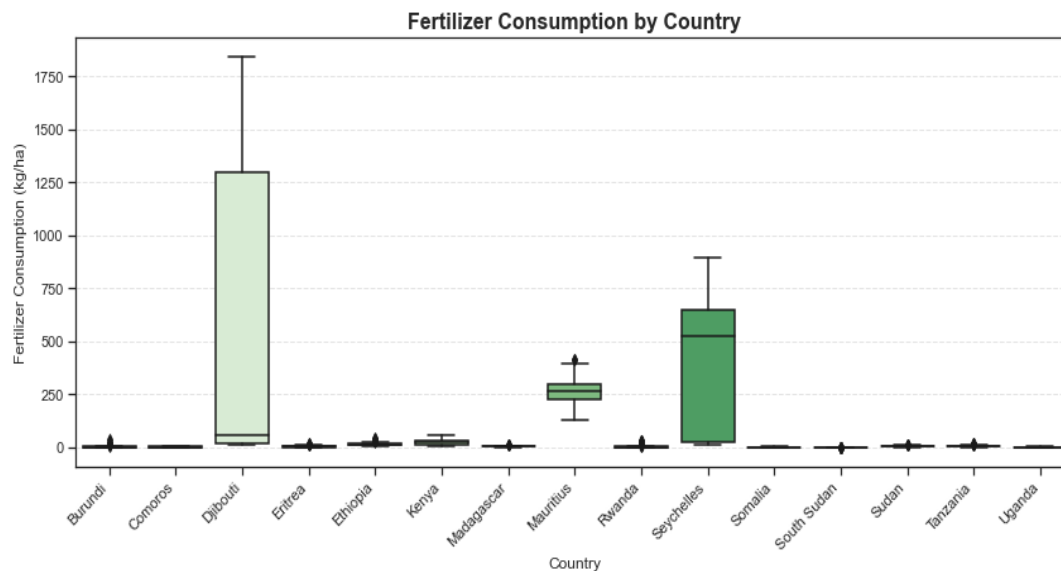
1. Fertilizer consumption (kg/hectare)



Observation:

- Most countries use relatively small amounts of fertilizer per hectare, while a smaller group uses much higher amounts.

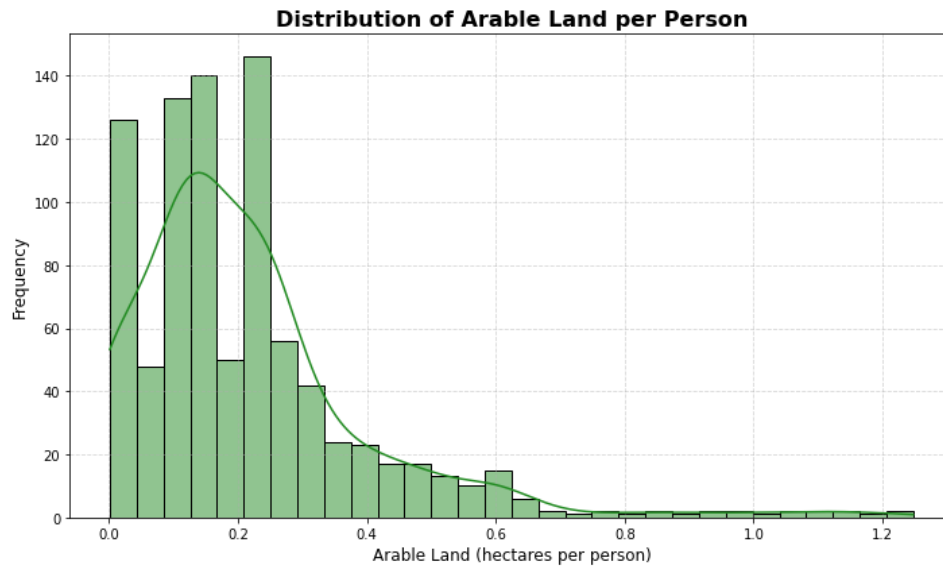
2. Fertilizer Consumption by Country



Observations:

- Djibouti and Seychelles are the top, in terms of fertilizer consumption in kgs per hectare.
- Mauritius follows at third, but the rest have relatively low levels of fertilizer consumption.

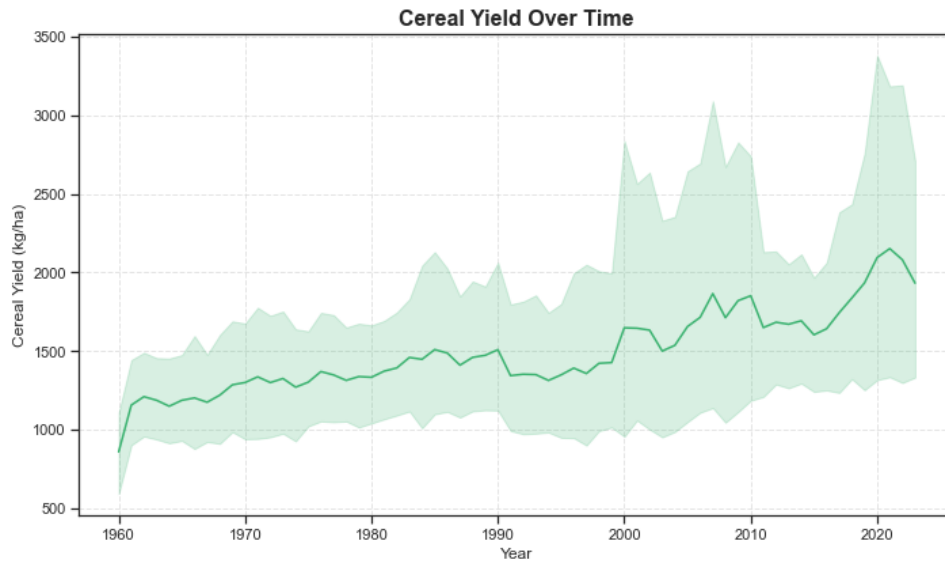
3. Distribution of Arable Land per Person



Observations:

- The majority of countries have very limited arable land available per person, with most falling below 0.3 hectares per individual.
- This indicates high population pressure on arable land in many regions, especially in more densely populated countries.

4. Cereal Yield Over Time



Observations:

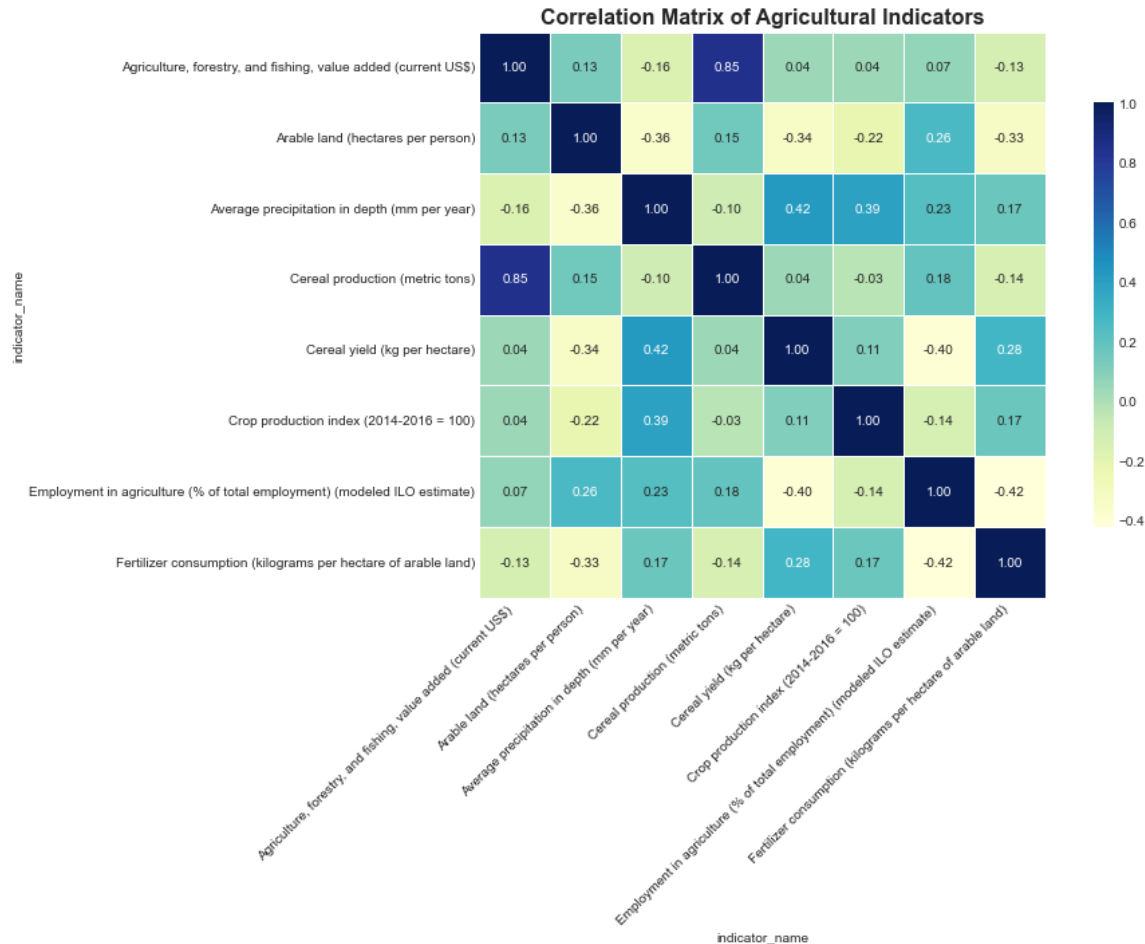
- Cereal yields have shown a clear upward trend globally since the 1960s, more than doubling in many cases. This reflects significant progress in agricultural practices, technology, and input use.
- However, the wide variation around the trend—especially in recent years—suggests that not all countries are benefiting equally from these advancements.
- External factors such as climate variability, policy changes, and regional conflicts may be driving these fluctuations and need to be carefully managed to sustain growth.

Key insights:

- **Fertilizer Use:** Significant increase in Kenya and Ethiopia from 2000 onward. Uganda and Burundi show stagnation.
- **Cereal Yields:** Gradual improvement across the board. Ethiopia and Rwanda exhibit the steepest growth.
- **Land Use:** Proportion of arable and agricultural land has remained relatively stable in most countries.

3.2 Bivariate Analysis

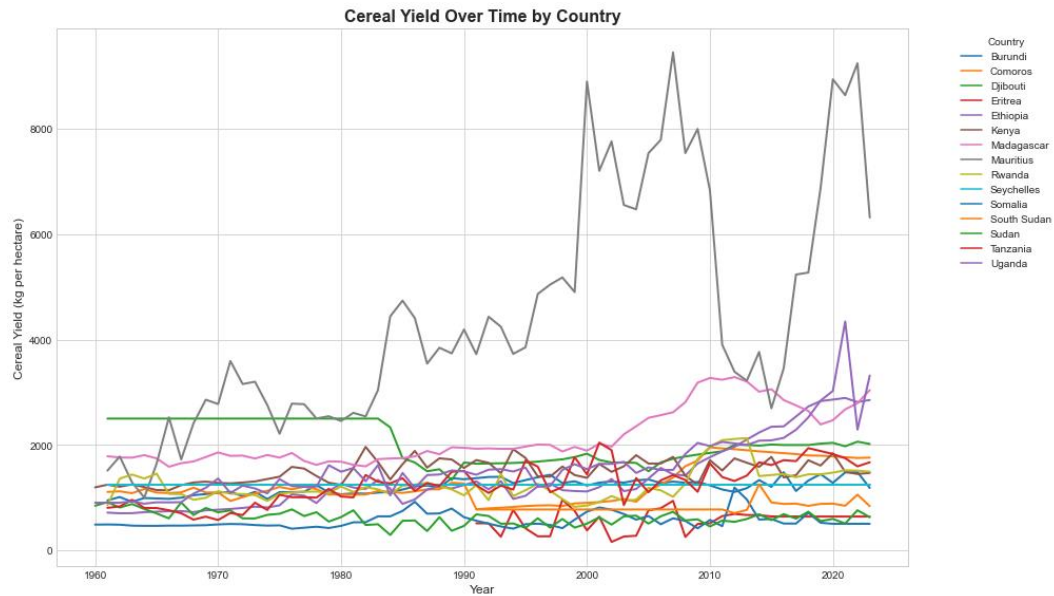
1. Correlation Matrix for Agricultural Indicators



Observations:

- Agricultural land and arable land are strongly correlated (0.85), meaning countries with more agricultural land per capita also tend to have more arable land per capita.
- Precipitation is positively correlated with cereal production (0.42) and renewable water (0.39), suggesting water availability supports agricultural output.
- Fertilizer consumption has weak or negative correlations with most variables, including cereal yield (-0.10), implying fertilizer use alone doesn't drive yield.
- Rural population and total population are negatively correlated with renewable water (-0.40), suggesting higher populations may strain water resources.

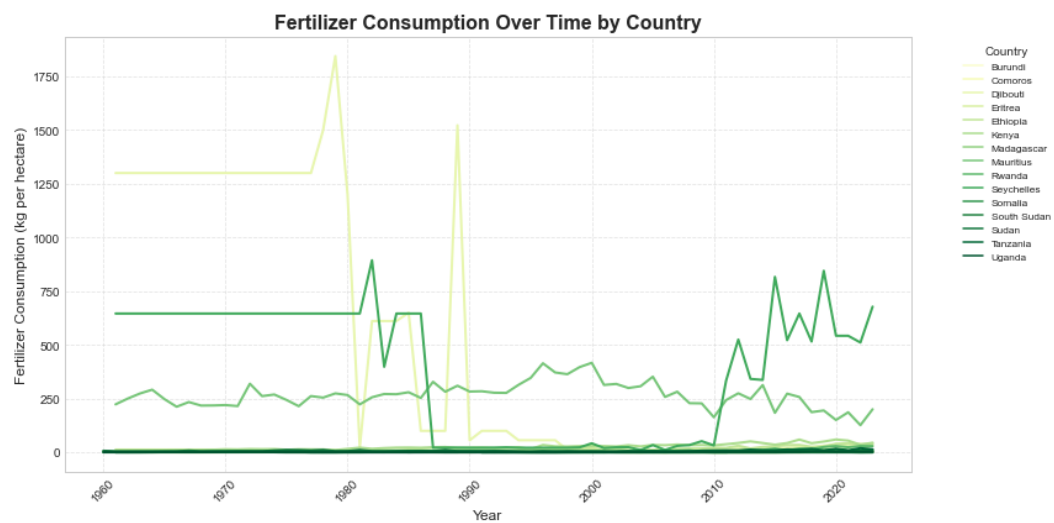
2. Cereal Yield Over Time by Country



Observations:

- There's a large disparity in productivity among countries.
- Yield improvement seems gradual for most, but a few (like Mauritius and Ethiopia) show more dramatic improvements.

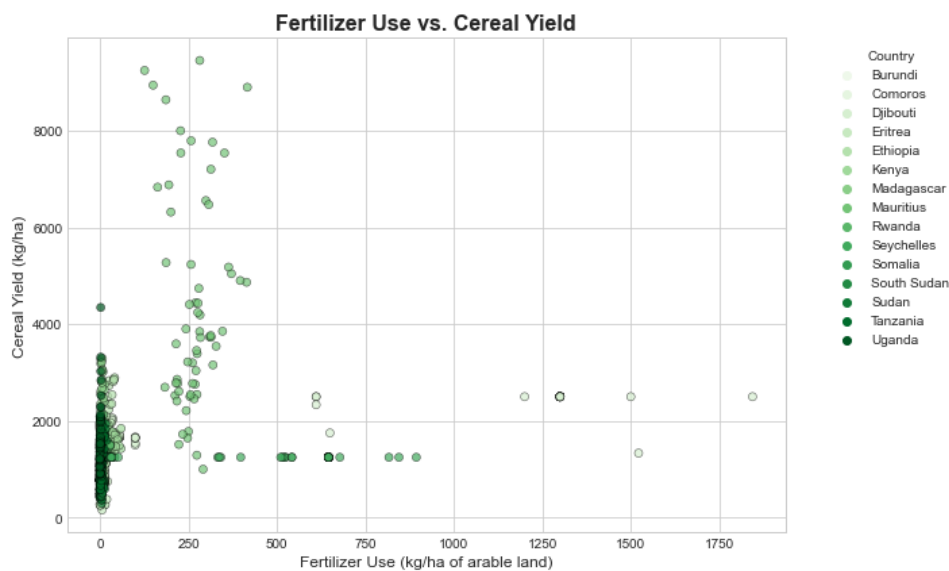
3. Fertilizer Consumption Over Time by Country



Observations:

- Seychelles and Mauritius have long-standing efficient agricultural systems.
- Kenya and Rwanda are emerging agricultural economies improving their fertilizer usage recently.
- Djibouti's sharp spikes and collapses show how unsustainable interventions fail to lead to lasting change.
- Most East African nations still struggle with low fertilizer adoption, which directly connects to lower cereal yields and food insecurity risks.

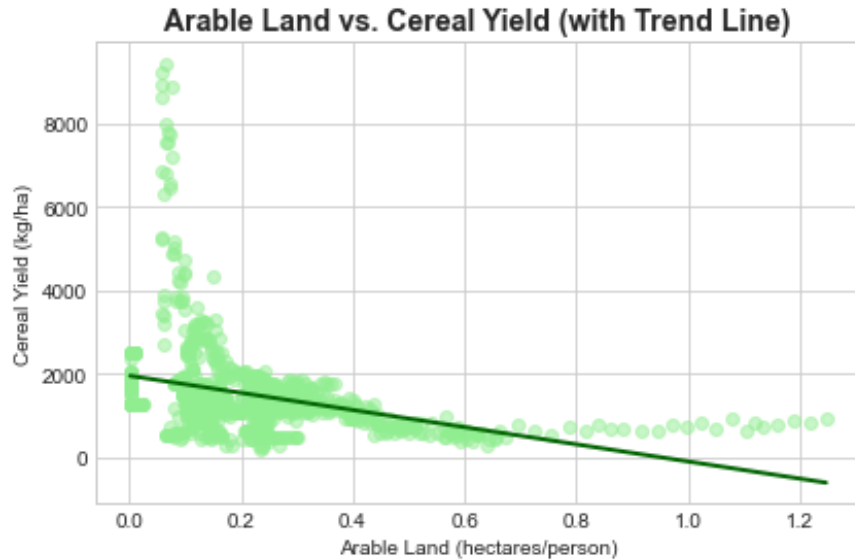
4. Fertilizer Use vs Cereal Yield



Observations:

- Most countries cluster at low fertilizer usage, but their yields vary widely.
- A few countries use high amounts of fertilizer, yet do not always achieve higher yields.

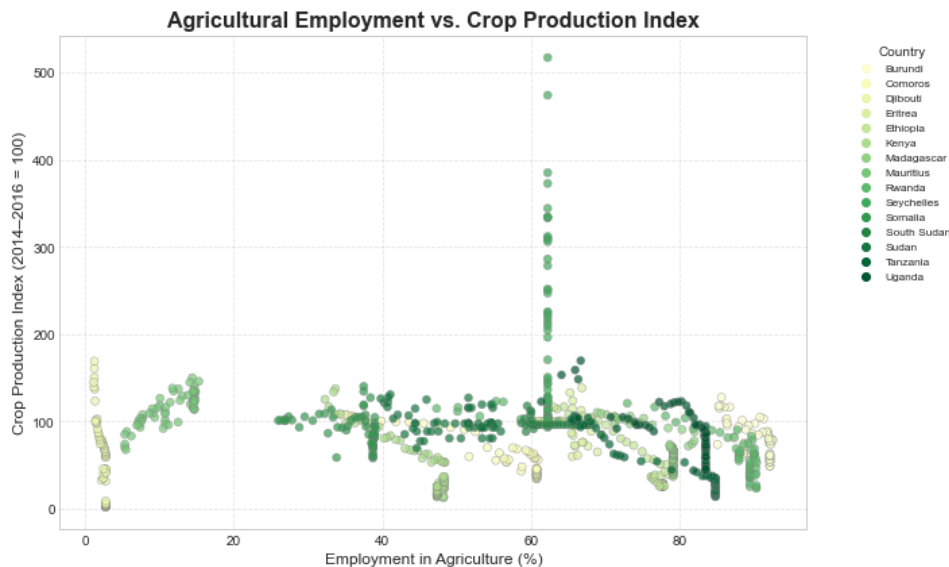
5. Arable Land per Person vs. Cereal Yield



Observations:

- **Negative Correlation:** Countries with more arable land per person generally have lower cereal yields per hectare, as shown by the downward trend line.
- **High Yield with Less Land:** Higher cereal yields are mostly seen where arable land per person is limited, likely due to intensive farming methods and better agricultural inputs.
- **Low Yield with More Land and Outliers:** Countries with more land per person often show lower yields, while a few outliers with very low land availability achieve exceptionally high yields, possibly from advanced farming systems or specialized crops.

6. Agricultural Employment vs. Crop Production Index



Observations:

- Countries with lower agricultural employment percentages often achieve higher crop production indices, indicating more efficient and modernized farming practices.
- In contrast, nations with higher agricultural employment (above 40–50%) generally see stagnant or lower crop productivity, suggesting reliance on labor-intensive, less efficient agricultural systems.
- A few outliers show very high crop production despite moderate employment levels, likely due to technological advancements or targeted agricultural improvements.

Key Insights:

- **Positive correlation between fertilizer use and cereal yield:** Countries with higher fertilizer input tend to produce more cereals per hectare.
- **Kenya and Ethiopia show stronger fertilizer-efficiency trends.**
- Heatmaps and scatter plots revealed:
 - **A moderate to strong positive relationship ($r \approx 0.6 - 0.75$)** between fertilizer consumption and cereal yield.
 - Weak relationship between agricultural land % and productivity, suggesting land alone is not the productivity driver.

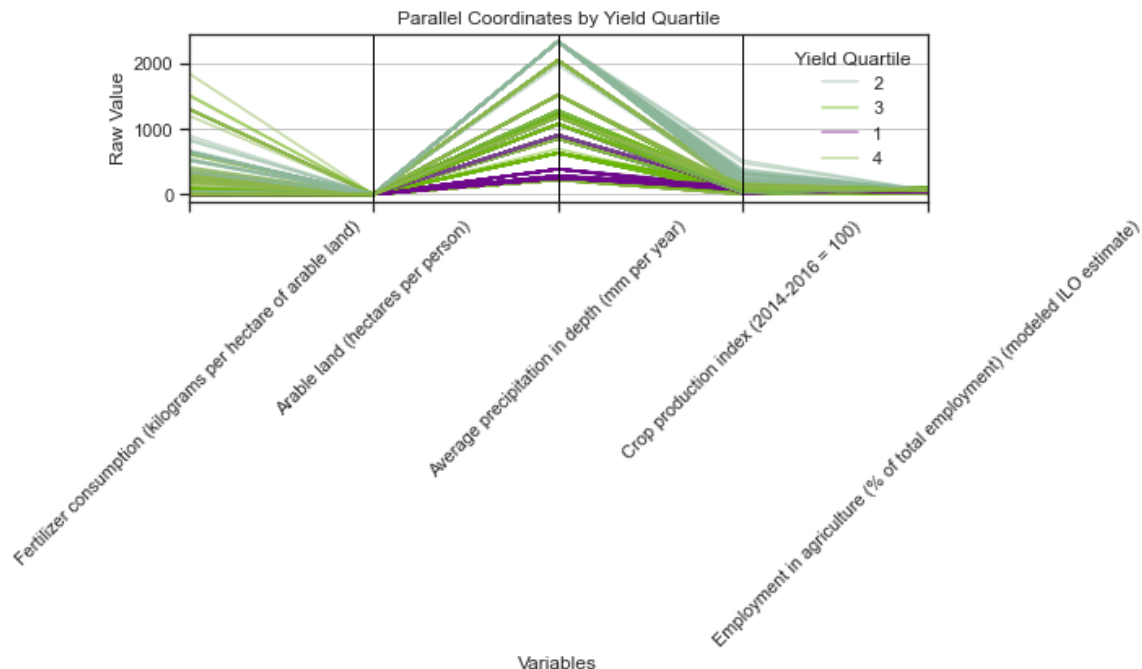
3.3 Multivariate Analysis

1. Pair Grid for All Variables

A Pair Grid visualization showed several important relationships: fertilizer consumption positively correlates with cereal yields, while higher agricultural employment tends to associate with lower fertilizer use, suggesting differences in farming intensification.

Meanwhile, land availability and precipitation patterns did not show strong direct relationships with productivity, highlighting that management practices and input quality are more decisive than natural endowments alone.

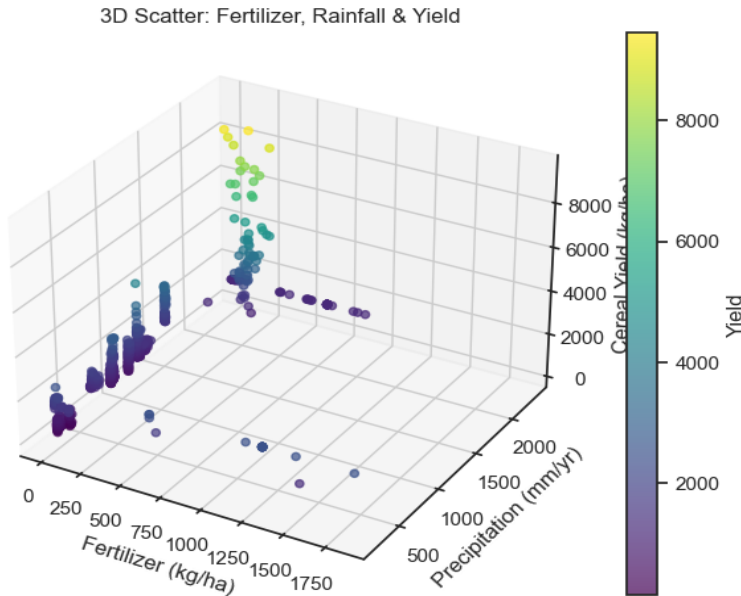
2. Parallel Coordinates plot



- Using a parallel coordinates plot grouped by cereal yield quartiles, it became clear that high-yielding countries (Q4) generally apply more fertilizer, achieve higher crop production indices, and have a lower share of their workforce in agriculture, implying more mechanized and efficient farming.
- Conversely, low-yield countries (Q1) showed less fertilizer use, lower production indices, and heavier dependence on manual labor, reinforcing the importance of agricultural modernization for boosting yields.

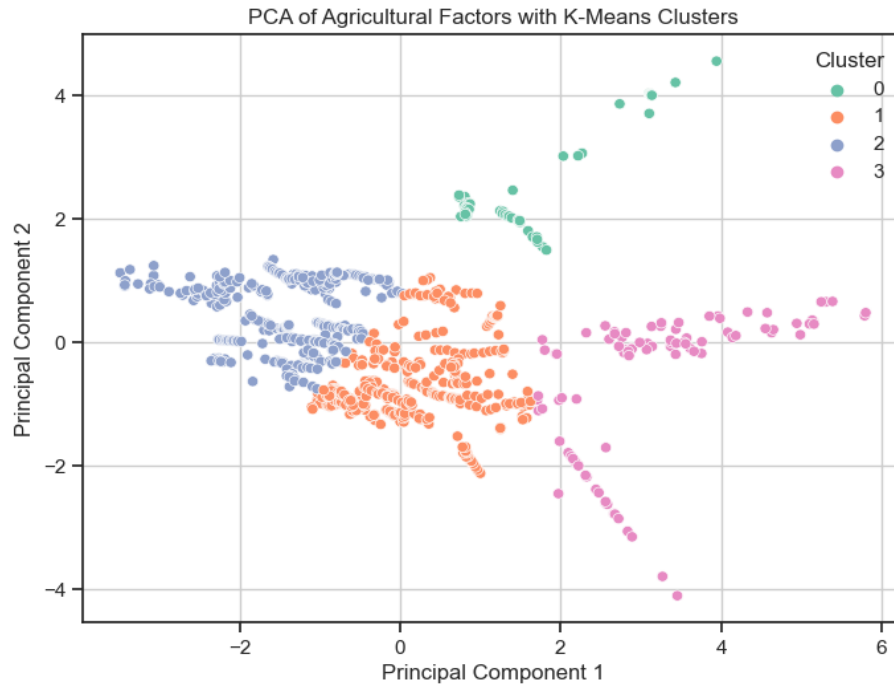
3. 3D Scatter Plot

- Further insights came from a **3D scatterplot** of fertilizer use, precipitation, and cereal yield, which showed that optimal yields occur with moderate-to-high fertilizer use (around 750–1250 kg/ha) and moderate rainfall (approximately 1000–1500 mm/year); extremely high or low values of either factor limit production.
- Correlation analysis confirmed these trends, highlighting fertilizer consumption as the strongest positive driver of cereal yield among the variables studied.



4. Principal Component Analysis

- Finally, **Principal Component Analysis (PCA)** was used to reduce dimensionality and visualize similarities between country-year observations. Most variation was captured along the first principal component (PC1), representing contrasts in input intensity and productivity.
- **K-Means clustering** applied to the PCA outputs identified four distinct groups of countries, which roughly correspond to varying levels of agricultural development and efficiency across the dataset.



- Clustering methods (e.g., K-Means) reveal **three key groups** of countries:
 1. **High-input, high-output:** Kenya, Ethiopia
 2. **Moderate-input, improving output:** Rwanda, Tanzania
 3. **Low-input, low-output:** Uganda, Burundi

4.0 Modeling

This chapter details the modeling efforts undertaken to forecast fertilizer consumption in Kenya and to predict the Crop Production Index based on agricultural input factors. Two separate modeling approaches were employed:

- **Time Series Forecasting** for fertilizer consumption
- **Regression Modeling** for crop production prediction.

4.1 Time Series Forecasting for Kenya's Fertilizer Consumption

4.1.1 Data Preparation

The analysis focused on Kenya's fertilizer consumption data from 1960 to 2023. The data was filtered to retain only Kenya-specific records, focusing on the "Fertilizer consumption (kilograms per hectare of arable land)" variable.

- The year column was converted into a datetime format to ensure proper chronological order.
- The data was resampled to an annual frequency to maintain consistency and allow for smoother forecasting.

4.1.2 Model Selection and Training

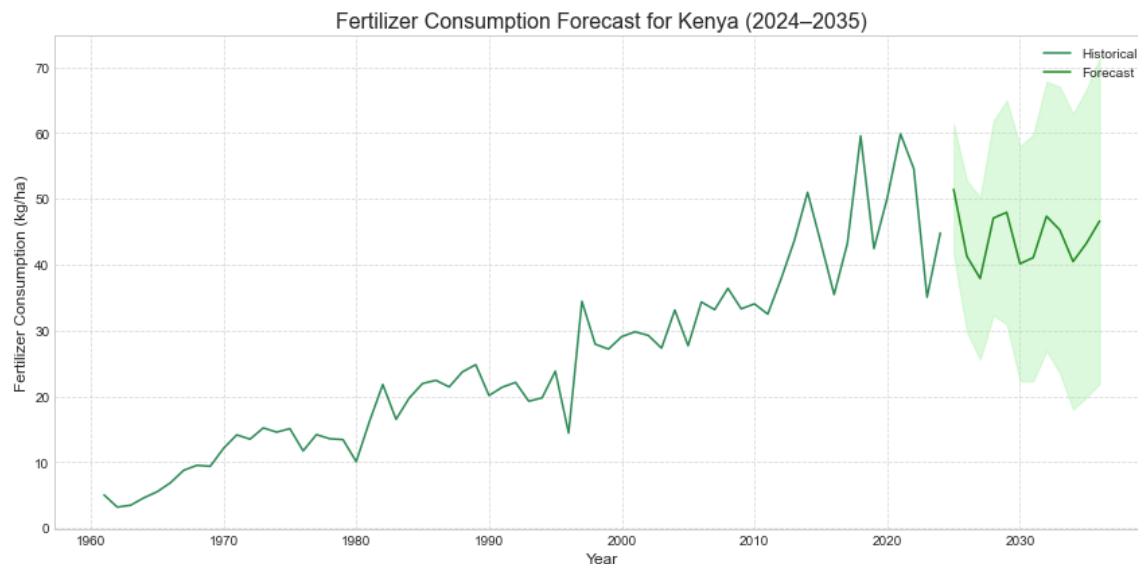
An **Auto-Regressive Integrated Moving Average (ARIMA)** model was selected for the time series forecasting. Specifically, an **ARIMA (2,1,2)** configuration was manually chosen based on initial experimentation to balance complexity and performance.

- **AR (2)**: Accounts for two past periods' values.
- **I (1)**: Indicates that the series is differenced once to achieve stationarity.
- **MA (2)**: Incorporates moving average terms from two previous forecast errors.

4.1.3 Forecasting Future Consumption (2024–2035)

- A forecast for the next **12 years (2024–2035)** was generated. Alongside the point estimates, 95% confidence intervals were also computed to capture the forecast's uncertainty.
- A future year index was created to plot these projections against historical data.

4.1.4 Visualization



The figure displays:

- **Historical fertilizer consumption** (1960–2023) in dark green
- **Forecasted fertilizer consumption** (2024–2035) in light green
- **Confidence intervals** shaded around the forecast to visualize prediction uncertainty

4.1.5 Interpretation of Forecast

- The historical data reveals a long-term upward trend in fertilizer consumption, characterized by periods of fluctuation and sharp increases, particularly noticeable over the last two decades.

The forecast suggests:

- Moderate stabilization around 40–45 kg per hectare over the next decade.
- A departure from previous periods of sharp increase, indicating a more gradual growth or plateau.

The wide confidence intervals highlight considerable uncertainty, suggesting that external factors such as:

- Government agricultural policies
- Global fertilizer prices
- Climate change impacts
- Adoption of new farming technologies

could significantly influence actual fertilizer consumption trends.

Insight:

- Strategic policy interventions can help drive fertilizer consumption toward the higher end of the forecast range, improving crop yields and food security.

4.2 Predicting Crop Production Index Using Regression Models

- In addition to fertilizer consumption forecasting, regression modeling was undertaken to predict the Crop Production Index (2014–2016 = 100) based on critical agricultural inputs.

4.2.1 Feature and Target Selection

The following variables were selected:

- **Target Variable:**
 - Crop Production Index (2014–2016 = 100)
- **Predictor Variables:**
 - Fertilizer consumption (kilograms per hectare of arable land)
 - Cereal yield (kg per hectare)
 - Arable land (hectares per person)

4.2.2 Handling Missing Values

Missing values in the features were handled using **mean imputation**, replacing missing entries with the respective column averages.

4.2.3 Data Splitting and Standardization

The dataset was split into **80% training data** and **20% testing data** to evaluate model generalization performance.

To ensure uniformity across features, **standardization** was applied.

4.2.4 Models Trained

Three different regression models were trained and evaluated:

Model	Description
Linear Regression	Baseline model assuming a linear relationship
Random Forest Regressor	Ensemble model combining multiple decision trees
XGBoost Regressor	Gradient boosting model optimized for speed and performance

4.2.5 Model Performance

The performance of the models was evaluated using:

- **Mean Squared Error (MSE):** Measures average squared difference between actual and predicted values (lower is better)
- **R-squared (R^2):** Measures proportion of variance explained by the model (higher is better)

Model	Mean Squared Error (MSE)	R-squared (R^2)
Linear Regression	1906.41	0.08
Random Forest Regressor	257.66	0.87
XGBoost Regressor	270.22	0.87

4.2.6 Interpretation of Results

- **Linear Regression** performed poorly with an R^2 value of just 0.08, indicating that a simple linear approach does not capture the underlying patterns between agricultural inputs and crop production effectively.
- **Random Forest Regressor** achieved the best performance with an R^2 of 0.875 and the lowest MSE, indicating strong predictive capability and good model fit.
- **XGBoost Regressor** also performed very well, closely trailing Random Forest, suggesting that ensemble-based, non-linear models are better suited for this problem.

4.2.7 Final Model Selection

Given the high R-squared score and low error rate, **Random Forest Regressor** was selected as the **best model** for predicting Kenya's Crop Production Index based on the available features.

Recommendations

1. **Expand Fertilizer Accessibility:** Establish regional supply chains and subsidy programs to make fertilizers more affordable and accessible to farmers.
2. **Promote Sustainable Fertilizer Usage:** Train farmers on correct fertilizer application techniques to improve efficiency and reduce environmental degradation.
3. **Enhance Cereal Crop Productivity:** Invest in research, better seed varieties, and irrigation infrastructure to boost cereal yields sustainably.
4. **Optimize Use of Arable Land:** Encourage land management practices such as crop rotation and soil conservation to maximize the productivity of available farmland.
5. **Strengthen Farmer Education Programs:** Launch region-wide agricultural extension services that provide farmers with up-to-date knowledge on inputs, yields, and market access.
6. **Leverage Technology and Innovation:** Promote adoption of precision farming tools, mobile-based advisory services, and AI-driven solutions to modernize agriculture.
7. **Build Resilience Against Agricultural Risks:** Develop regional strategies including crop insurance schemes and emergency fertilizer reserves to protect against climate and market shocks.
8. **Improve Agricultural Data Infrastructure:** Establish harmonized systems for real-time agricultural data collection, analysis, and sharing across East African countries.
9. **Encourage Public-Private Partnerships (PPPs):** Foster collaboration between governments, private companies, and NGOs to drive innovation, funding, and scaling of agricultural solutions.

Conclusion

This project highlights critical insights into fertilizer consumption patterns, land use, and crop production trends across East Africa. Our analysis reveals that while fertilizer use is gradually increasing, there is significant room for improvement in optimizing agricultural inputs and maximizing crop yields. Forecasting suggests a moderate rise in fertilizer consumption, but variability remains a concern due to external economic and environmental factors. Furthermore, advanced modeling shows that non-linear relationships between fertilizer use, cereal yield, and land availability strongly influence agricultural productivity, emphasizing the need for integrated solutions. Moving forward, a regional focus on accessibility, sustainability, innovation, and data-driven decision-making will be essential to unlock East Africa's full agricultural potential and ensure food security for the future.