

DATA REPORT

Optimizing Sourcing and Pricing Strategies for GlobeHarvest Co. in Kenya



Organization: Moringa School

Student pace: DS-full time-Online

Group members:

- Joyce Chepng'eno
- Sarah Karanja
- Brandon Muraya
- Pete Njagi
- Tom Mwabire
- James Koli

Date of publication: 10/05/2024

Overview

GlobeHarvest Co. is an emerging import-export company poised to enter the Kenyan agricultural market. With a commitment to leveraging data-driven strategies, GlobeHarvest aims to optimize its sourcing and pricing strategies to establish a competitive presence in Kenya's agricultural sector. By harnessing the power of data analytics, GlobeHarvest seeks to gain valuable insights into market dynamics, consumer preferences, and pricing trends to drive informed decision-making and maximize profitability.

Business Understanding

Kenya's agricultural market offers immense potential for GlobeHarvest Co., characterized by diverse regions with varying agricultural capabilities and consumer demands. However, navigating this complex landscape presents several challenges, including identifying cost-effective sourcing locations, strategically timing purchases, and ensuring competitive pricing amid price variabilities across regions. To succeed in the Kenyan market, GlobeHarvest must develop a deep understanding of market dynamics, consumer behavior, and competitive factors to optimize its procurement processes and pricing strategies effectively.

Challenges

- Identifying optimal sourcing locations.
- Timing purchases strategically.
- Navigating price variabilities.
- Ensuring Competitiveness by effective and accurate price prediction.

Stakeholder Engagement

GlobeHarvest Co. stakeholders include:

- Management team
- Data analysts
- Procurement team
- Sales and marketing team

Research Questions

- What are the cost-effective sourcing locations for various food commodities in different Kenyan counties?
- How can we strategically time purchases considering seasonal fluctuations and market dynamics?
- What strategies can be implemented to navigate price variabilities across different regions in Kenya?
- How do Kenyan prices compare to global averages, and what export opportunities does this present?

Problem Statement

GlobeHarvest Co. encounters obstacles in establishing a foothold in the Kenyan agricultural sector, necessitating solutions for efficient sourcing, strategic purchasing, price volatility management, and maintaining international competitiveness.

Main Objective

The primary objectives of GlobeHarvest Co. in optimizing its sourcing and pricing strategies for the Kenyan market.

Specific Objectives

- Identify cost-effective sourcing locations.
- Strategically time purchases.
- Navigate price variabilities.
- Ensure global competitiveness.

Data Understanding

Data Source

The dataset obtained from Kaggle consists of food price data collected in Kenya, providing valuable insights into the pricing dynamics of various food commodities across different regions and markets.

Dataset Overview

The dataset contains 10,767 rows, each representing a unique observation of food prices in a specific market on a particular date. This dataset provides a comprehensive view of food pricing trends and variations across different regions and market types in Kenya, offering valuable insights for analysis and decision-making in the agricultural sector.

The dataset the following columns:

- **date**: The date of the price observation.
- **admin1**: The first-level administrative division (e.g., province or county) in Kenya.
- **admin2**: The second-level administrative division (e.g., district or sub-county) in Kenya.
- **market**: The name of the market where the price was recorded.
- **latitude**: The latitude coordinates of the market location.
- **longitude**: The longitude coordinates of the market location.
- **category**: The category of the food commodity (e.g., cereals and tubers, pulses and nuts, meat, fish, and eggs).
- **commodity**: The specific name of the food commodity.
- **unit**: The unit of measurement for the price (e.g., KG for kilograms, G for grams).
- **priceflag**: Flag indicating the nature of the price observation (e.g., actual, aggregate).
- **pricetype**: The type of price (e.g., Wholesale, Retail).
- **currency**: The currency used for pricing (e.g., KES for Kenyan Shilling).
- **price**: The price of the commodity in the local currency.
- **usdprice**: The price of the commodity converted to USD.

Data Files

Our dataset contained the csv file with Kenyan food prices namely:

- wfp_food_prices_ken.csv

Data Preparation

The dataset was loaded using the Jupyter Notebook, where it was previewed for a better understanding of the columns and the relationships that exist between them.

Data Cleaning

Missing Values and Duplicates

The data frame was checked for missing values and duplicates but none were found.

Dropping columns and rows.

The columns 'priceflag' and 'usdprice' were dropped as they were deemed unnecessary for the analysis.

The first row contained duplicated information about the columns so it was removed.

Date column modifications

The date column was converted to datetime format for ease of analysis while columns for year and month were added.

Metric standardization

Since most of the commodities came in different units i.e. KG, ML, L, Unit and Bunch, the metrics were standardized to enable comparative studies.

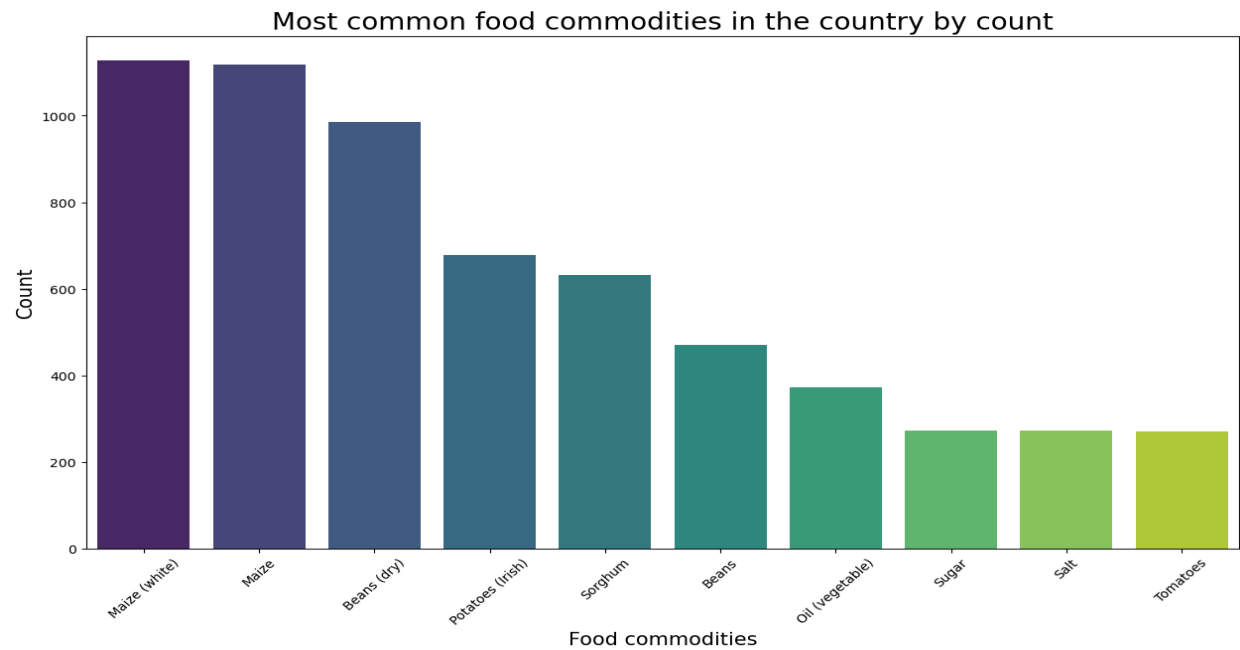
Column Modifications

The columns admin1 and admin2 were renamed to Provinces and Counties respectively.

Exploratory Data Analysis

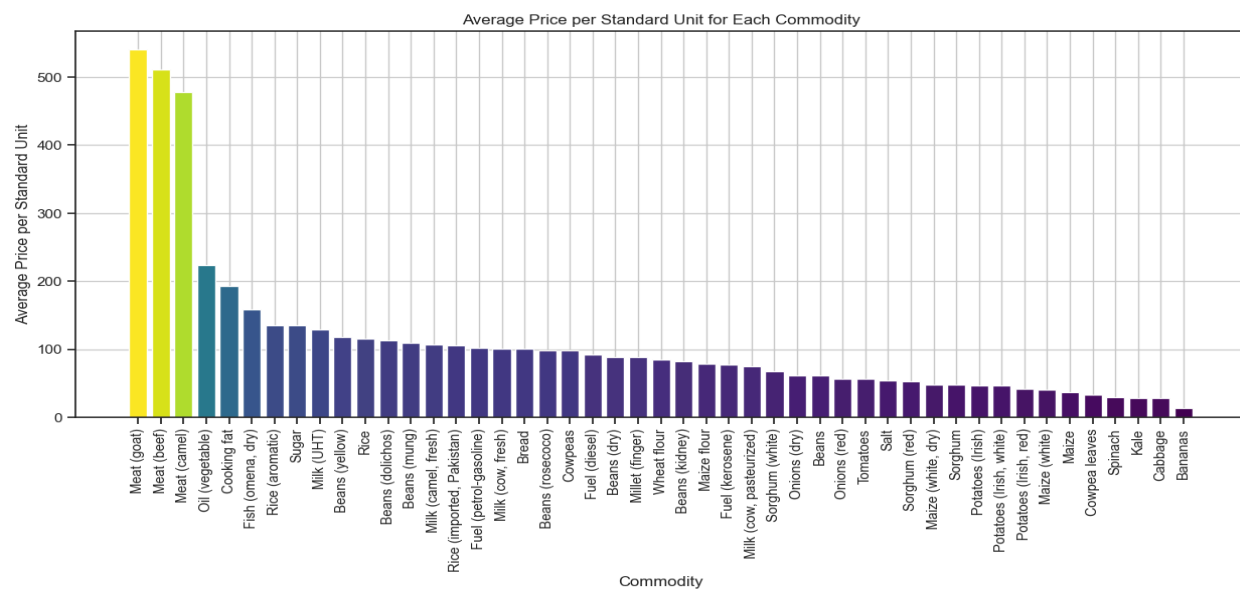
Most Common Food Commodities and Categories

Visualizations were generated to show the top 10 most common food commodities and categories in the dataset by count.



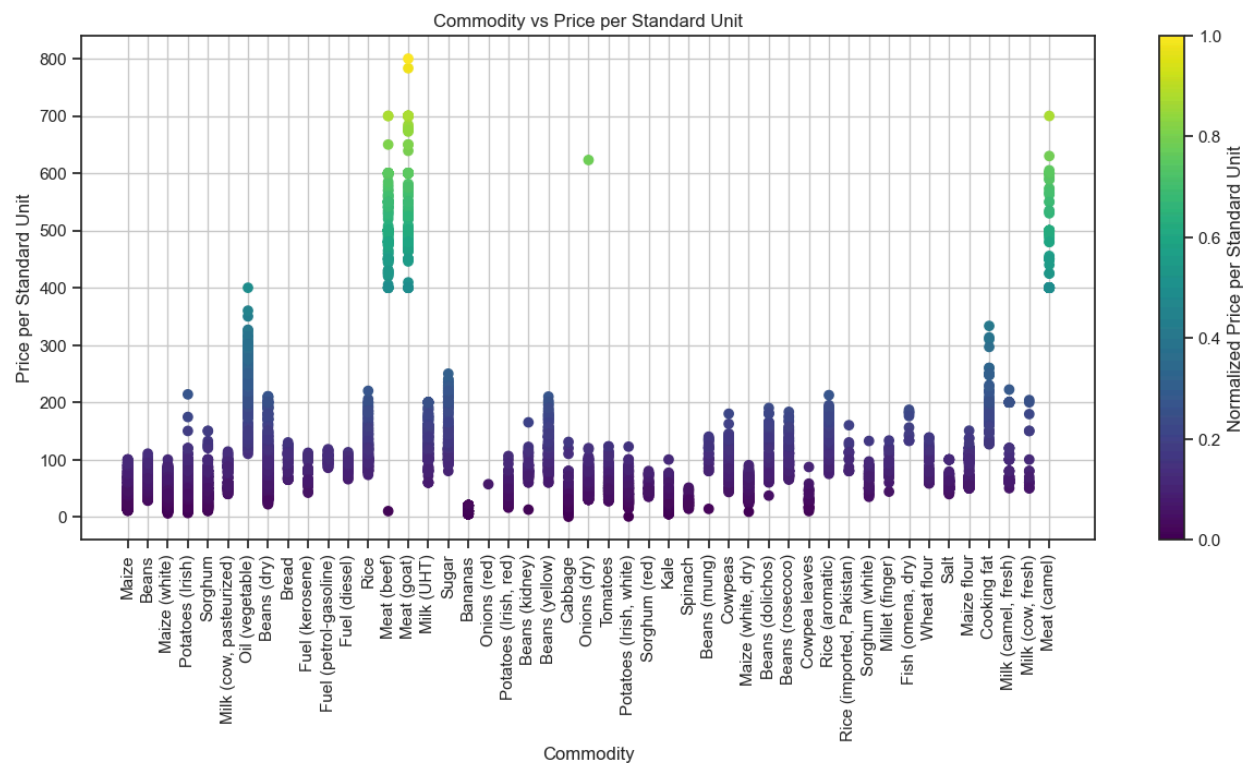
They showed that both varieties of maize seem to be the most popular followed by dried beans and potatoes(Irish).

Average Price per Standard Unit for Each Commodity



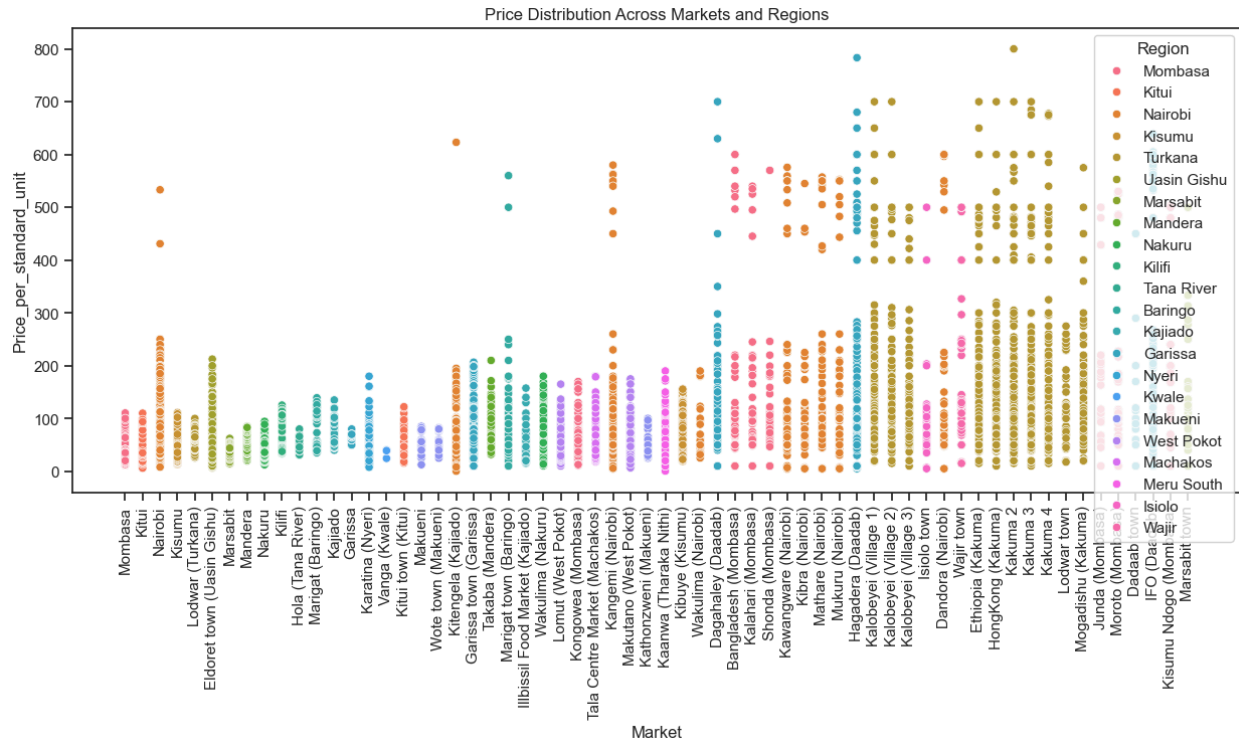
The visualization above shows the average price per standard unit of commodities in Kenya. From the visualization, Meat (goat, beef and camel) is the most expensive commodity. All the types of meat have an average price per standard unit of around 500 ksh. This is then followed by Vegetable Oil and cooking fat. The least expensive on the other hand are fruits and vegetables such as Bananas, Cabbage, Kale and Spinach.

Price Distribution Across Markets and Regions



The visualization above shows the price ranges of some of the top commodities across the country. For instance if we take a look at goat meat, in some areas in the country it could go for as low as 400ksh per kilo and in some areas in the country it could go for as high as 800ksh. This may be due to underlying factors such as ease of availability in the regions the price values were recorded from.

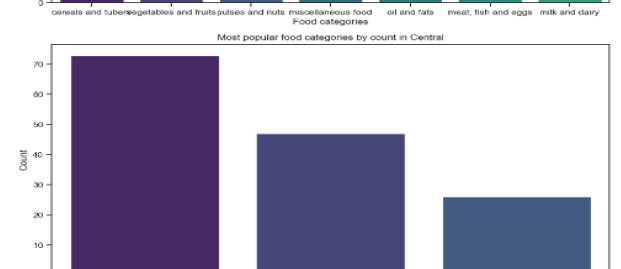
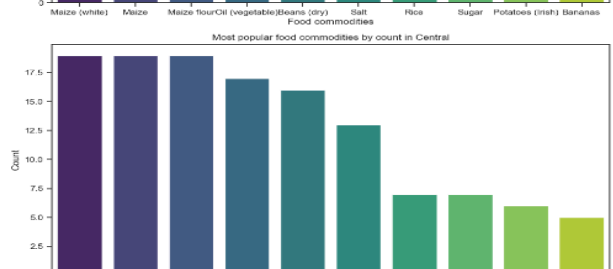
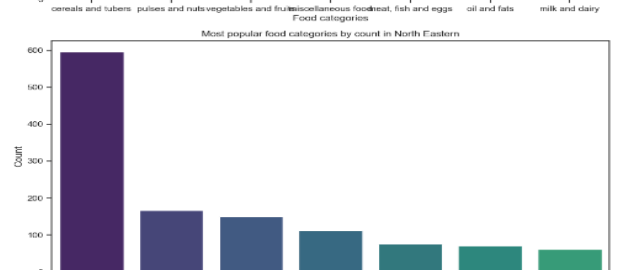
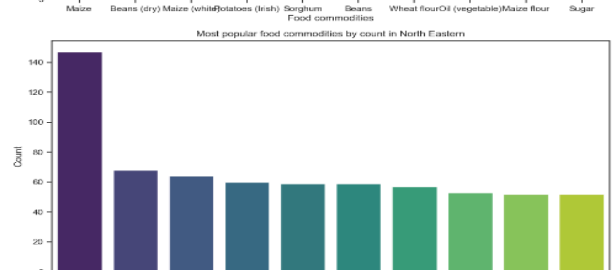
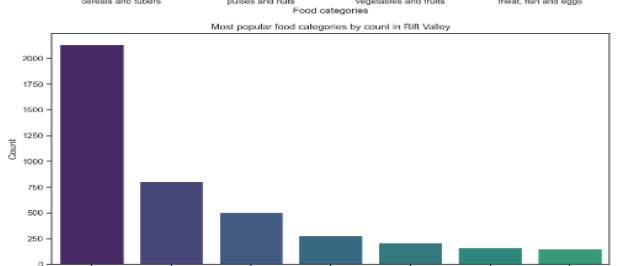
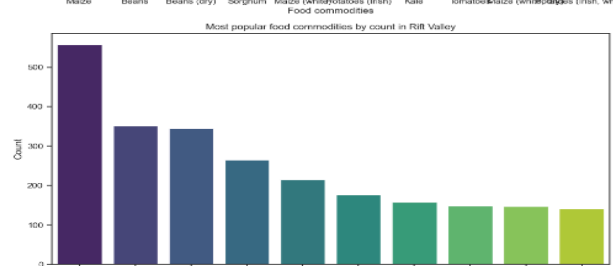
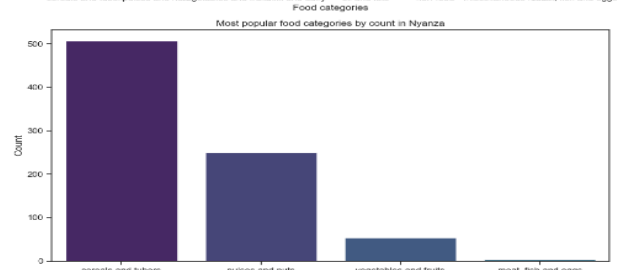
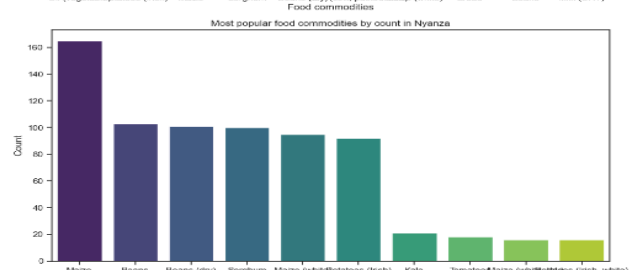
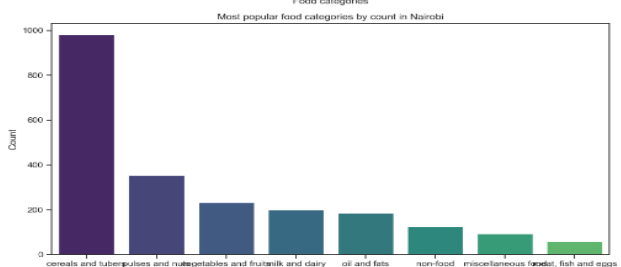
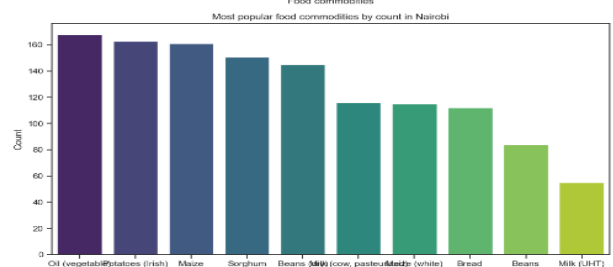
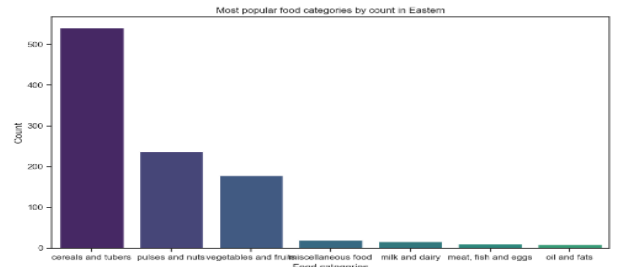
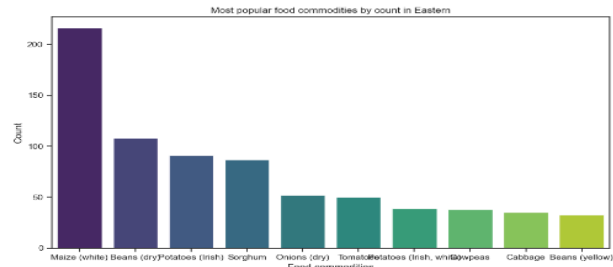
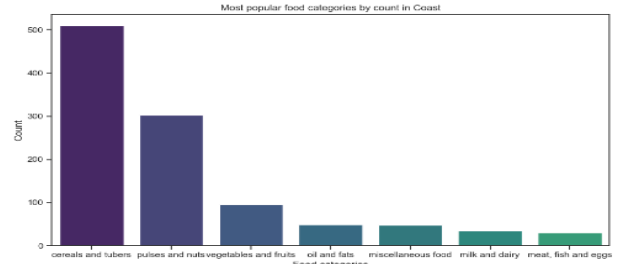
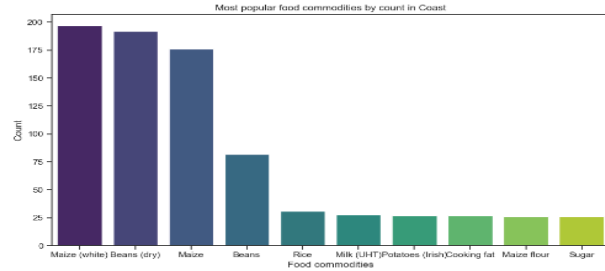
The other graph below shows the price distribution across selected markets and regions which are color coded depending on the counties they are in.



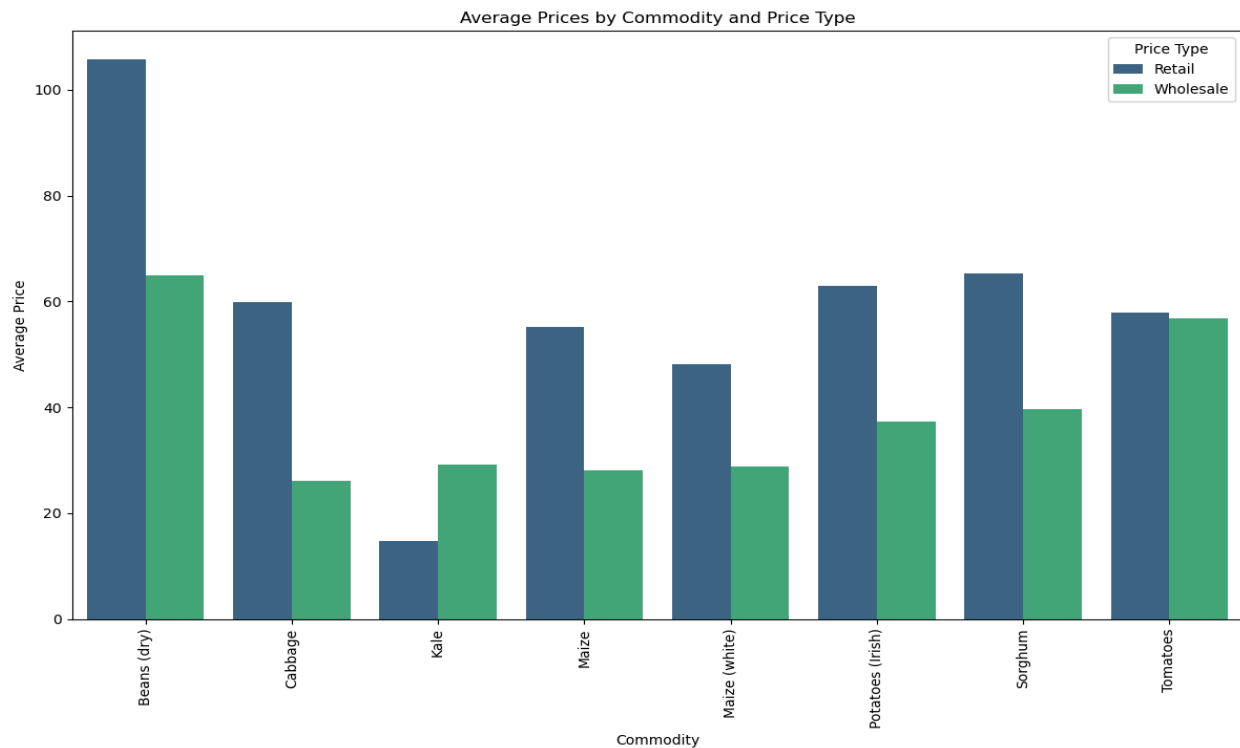
Most Popular Food Categories in Each Province

The visualizations below show the top 10 most popular food commodities and categories in the various Provinces across the country. This offers unique insight as to what residents of a certain region like consuming the most. For example;

- Central Province consumes Irish Potatoes the most while they mostly consume vegetables and fruits category of foods then followed by cereals and tubers.
- All other Provinces have maize as the most consumed food commodity.
- Strangely in Nairobi, Vegetable Oil is the most popular food commodity and is followed closely by Maize, Potatoes and Sorghum. This may be because a wider variety of food is consumed and the main constant ingredient in food preparation is vegetable oil. The most popular food category being cereals and tubers still backs up this theory.



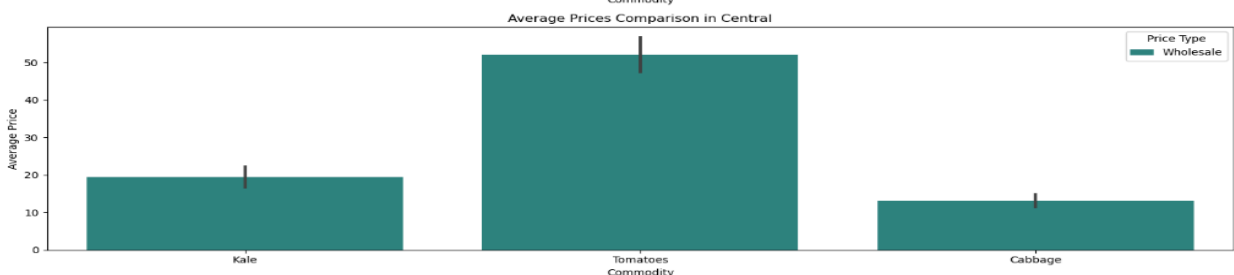
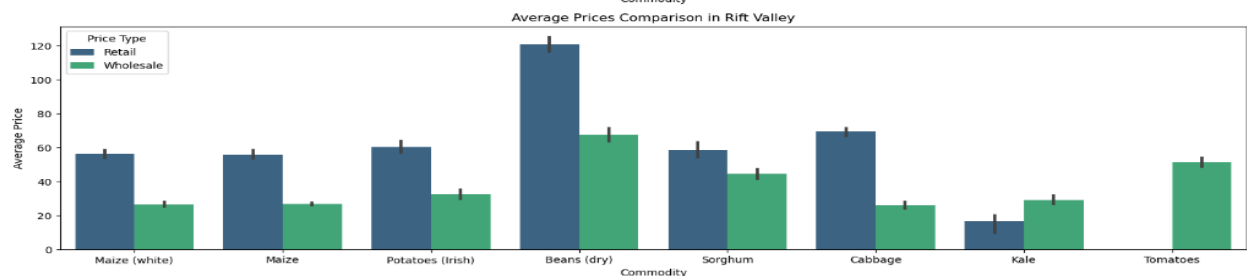
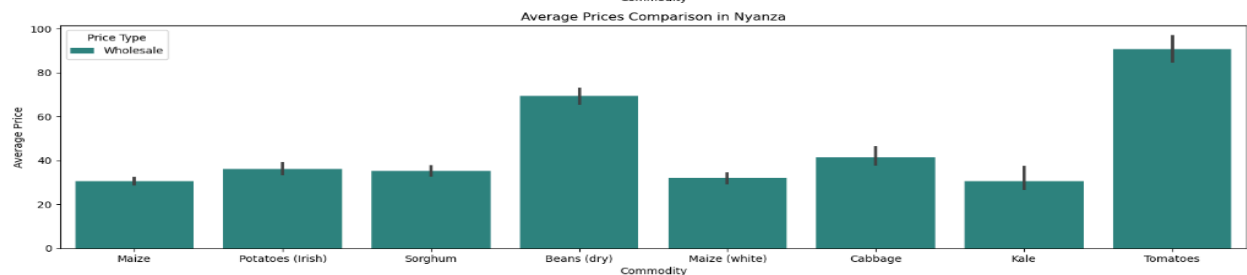
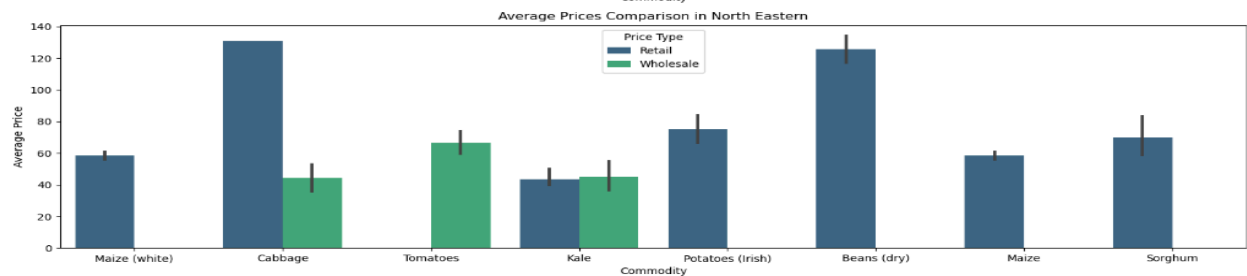
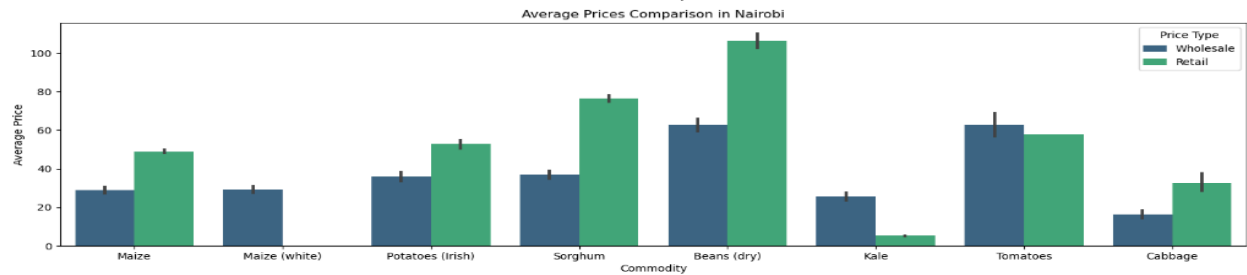
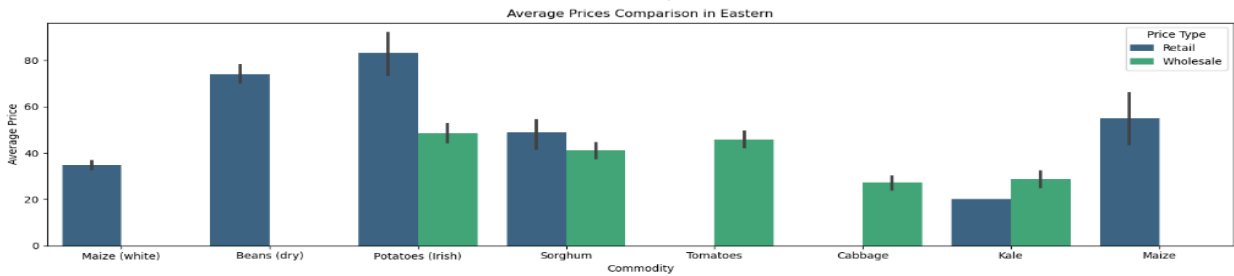
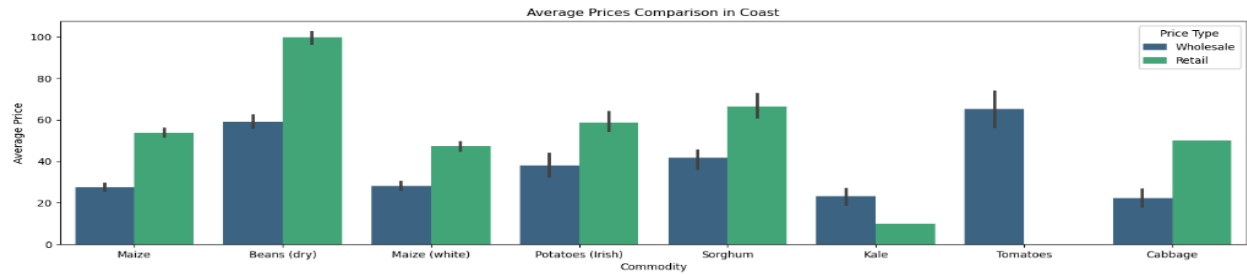
Comparing wholesale and retail prices



The plot above shows the average price per standard unit difference for commodities that are sold in both wholesale and retail. In almost all but Kale, the retail prices are more expensive than the wholesale prices. The biggest price difference is in Beans(dry) and Cabbage.

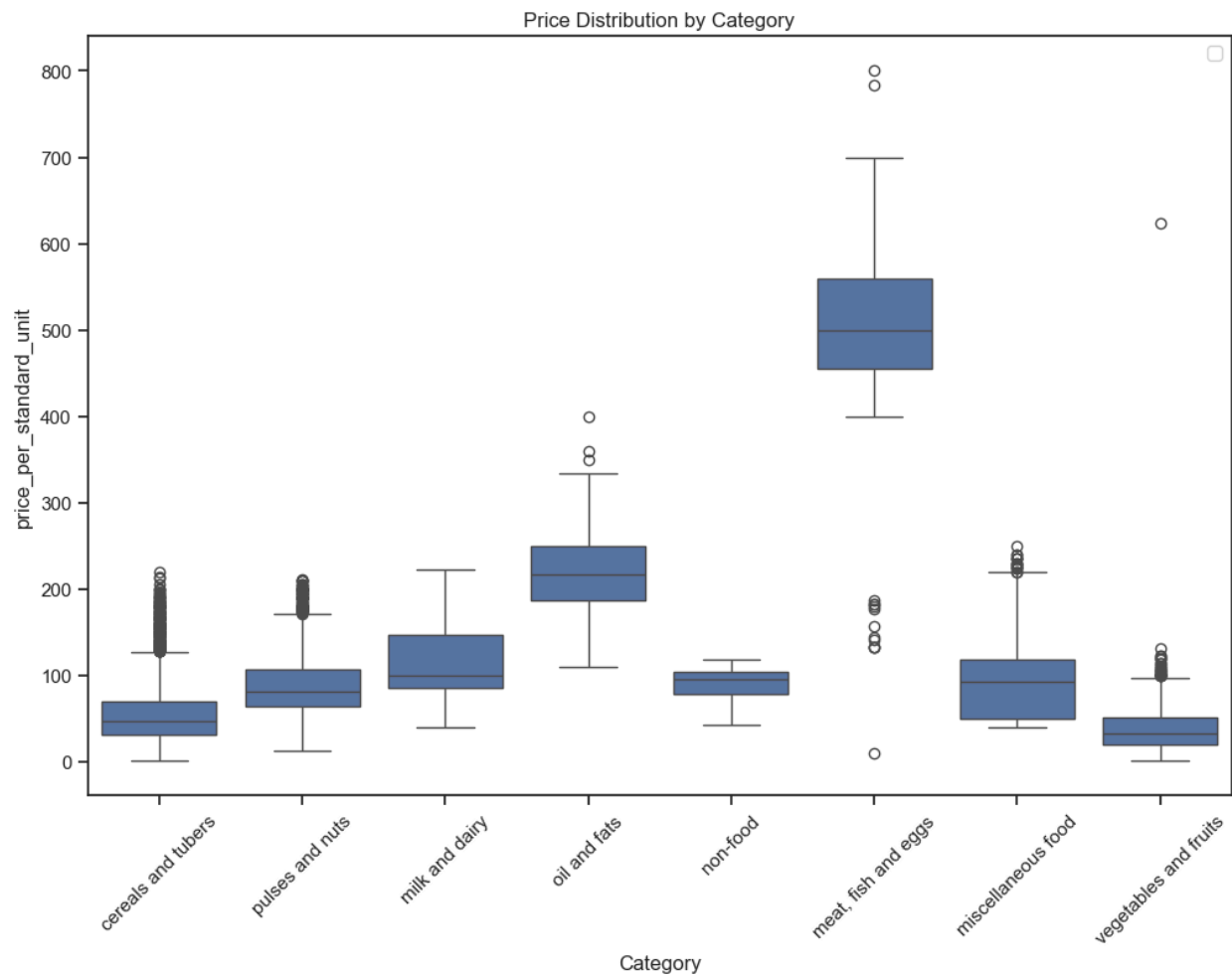
Average Prices Comparison Across Provinces

The visualizations below show the price per standard unit differences in commodities sold in retail and wholesale in Provinces across the country. For some commodities, the prices available are only either Retail or wholesale. Some Provinces (Central and Nyanza) have only one price type which is wholesale. This means that they tend to either buy only in wholesale or information on retail prices are not available.



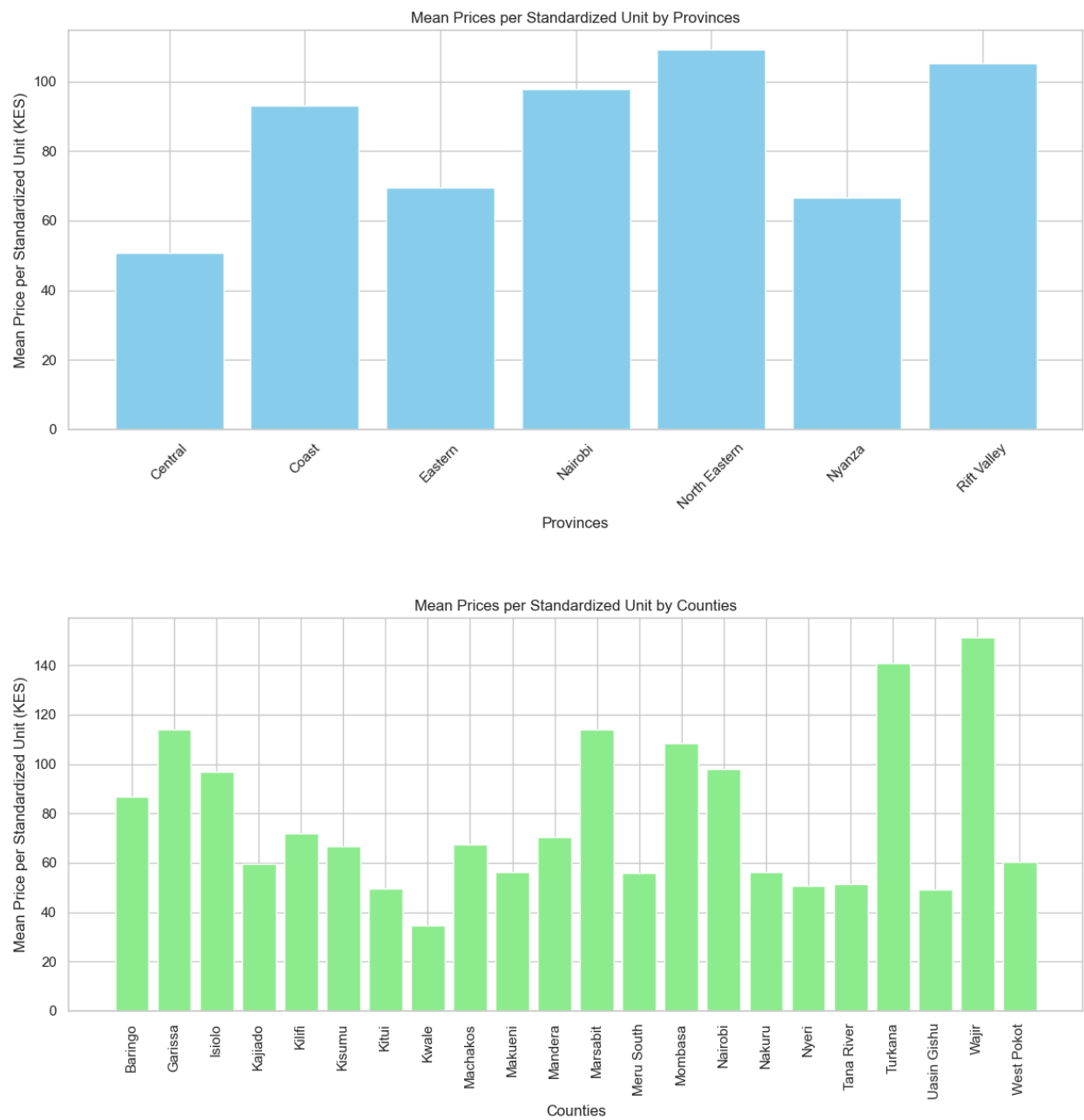
Price Distribution by Category

The visualization below shows the frequency distribution of the price per standard unit of all products and then breaks it down further to show the average distribution on average for each food category. From it, we can clearly see that the meat, fish and eggs category has on average the highest price per standard unit. It also has a bigger range to account for items like meat(camel) which is the highest meat price and fish(omena,dry) which is the lowest.



Market Segmentation

Mean Prices per Standardized Unit by Provinces and Counties



The visualizations above show the mean prices per standard unit for commodities in counties and Provinces. It can be noted that;

- Central Province has on average the lowest average price per standard unit followed by Eastern and Nyanza Provinces. North Eastern Province has on average the highest average price per standard unit.
- Kwale County has the lowest average price per standard unit. Wajir and Turkana have on average the highest prices per standard unit.

Mean Prices per Standardized Unit by Price Type



The visualization above shows that mean prices per standard unit for both Retail and Wholesale. This visualization shows that on average the Wholesale price is half the Retail price.

Overall, the EDA provided valuable insights into the distribution, pricing, and popularity of food commodities and categories across different regions and markets in Kenya. These findings will inform further analysis and decision-making processes for GlobeHarvest Co. as it seeks to optimize its sourcing and pricing strategies in the Kenyan agricultural market.

Modeling

Model Selection

In this section, various machine learning models are evaluated and compared to select the most suitable one for predicting future prices of wholesale commodities.

Linear Regression Model

The linear regression model was the first one attempted. Here are the results:

Mean Squared Error (MSE): 202.87 Coefficient of Determination (R^2 Score): 0.7943

Random Forest Regressor

Next, a Random Forest Regressor was trained and evaluated:

Mean Squared Error (MSE): 96.62 Coefficient of Determination (R^2 Score): 0.9020

Gradient Boosting Regressor

A Gradient Boosting Regressor was also tested:

Mean Squared Error (MSE): 96.62 Coefficient of Determination (R^2 Score): 0.9020

Decision Tree Regressor

Lastly, a Decision Tree Regressor was included:

Mean Squared Error (MSE): 211.58 Coefficient of Determination (R^2 Score): 0.7855

Model Evaluation

Comparing the models, the Random Forest Regressor outperformed the others in terms of both MSE and R^2 Score.

Random Forest Regressor: MSE = 96.62, R^2 = 0.9020

Gradient Boosting Regressor: MSE = 96.62, R^2 = 0.9020

Linear Regression: MSE = 202.87, $R^2 = 0.7943$

Decision Tree Regressor: MSE = 211.58, $R^2 = 0.7855$

Therefore, the Random Forest Regressor was chosen as the final model.

Hyperparameter Tuning

To optimize the Random Forest model further, a grid search was performed to find the best hyperparameters:

Our best parameters were found to be:

```
{'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 100}
```

Final Model

After tuning, the Random Forest Regressor with the best hyperparameters was evaluated on the test set:

Mean Squared Error (MSE): 27.31 Coefficient of Determination (R^2 Score): 0.9723

The significantly reduced MSE and improved R^2 score demonstrate the effectiveness of the tuned Random Forest model in predicting future wholesale commodity prices.

Conclusion

Some of the takeaways from the project are;

- **Cereals and Tubers Dominate Consumption:** EDA revealed that cereals and tubers are the most consumed food category nationwide. Global Harvest Co. should spend more on acquiring and selling cereals and tubers. This is because there is the highest demand for this food category.
- **Strategic Sourcing and Selling Locations:** From market cluster analysis, provinces with lower wholesale prices are ideal for sourcing while provinces with higher retail prices offer better selling opportunities. GlobalHarvest Co. should also consider selling retail in Provinces like Rift Valley, North Eastern and Nairobi. This is because they have the 3 highest retail prices per standard unit and will offer more in terms of profit.

- Targeting Specific Markets: By clustering markets based on average price per standard unit; markets in cluster 1 offer the lowest prices.
- Localized Sourcing Recommendations: GlobalHarvest Co. should put considerable effort in sourcing from Provinces that have lower Wholesale average prices per standard unit. In this instance Coast would be recommended then Easter and Rift Valley.
- Effective Price Prediction Model: Random Forest Regressor proved highly accurate(97% accuracy after tuning). Global Harvest Co. can use it to forecast prices effectively, enabling better procurement and pricing strategies for decision-making.

Recommendations

The following are the recommendations to be made after undertaking this project. This will also include ways to improve the system in future and general observations. These include;

- Global Harvest Co. should prioritize investments in acquiring and selling cereals and tubers, capitalizing on their high-demand
- Focus on provinces like Coast for sourcing and Rift Valley, North Eastern, and Nairobi for retail sales could optimize profits.
- Global Harvest Co. should consider sourcing from cluster 1 to minimize procurement costs and enhance competitiveness.
- Global Harvest Co. can forecast prices using the Random Forest Regressor model, enabling better procurement and pricing strategies.

Future Improvement Ideas

These include ways to improve the system in future and general observations. These include;

- Complete data can be acquired from other trusted sources so as to make the information more complete. This will improve prediction accuracy.
- More recent data can be acquired from the World Food Program since this data is updated monthly. This could be via api integration so as to make this process automatic.

Deployment

The deployment of the machine learning model on the web application enables users to obtain predictions on food categories based on descriptions and year of creation. The application

provides a convenient and accessible platform for users to utilize the predictive capabilities of the model. Continuous monitoring and updates will be conducted to ensure the optimal performance and reliability of the deployed system.