Global Rice Production Analysis

Hari Prasannaa Thangavel Ravi[1], Jin Hyuk Son[2], & Pranjal S. Wakpaijan[3]

[1,2,3] The George Washington University

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

Abstract

Rice production is vital for global food security and sustainable agriculture. This research project, titled "Global Rice Production Analysis," aims to explore the foundational stages of data-driven research in this context. Using the dataset "Explore Data on Agricultural Production," which comprises about 13,500 rows and 40 columns. This study systematically examines key factors influencing rice yield, including land use efficiency, production per capita, import and export dynamics, waste generation, and domestic supply. The study begins with cleaning the dataset, addressing missing values and outliers, followed by examining data distribution and interrelationships between variables. The project formulates and answers three SMART questions, employing exploratory data analysis (EDA) to visualize the results. Various statistical tests were conducted to validate these findings. This study not only addresses practical challenges in rice cultivation but also provides insights to refine agricultural strategies, optimize resource allocation, and promote sustainable practices. The findings aim to assist rice producers in enhancing yield predictions and contribute to strengthening global food security.

Keywords: Land use efficiency, Production per capita, Outliers, Data cleaning, Variable interrelationships

Word count: 162

Global Rice Production Analysis

Introduction

Rice is a staple food for over half of the global population, making its production essential for food security worldwide. As global demand for rice continues to rise due to population growth and evolving dietary habits, understanding the dynamics of rice production, trade, and sustainability becomes increasingly vital. Sustainable agricultural practices not only aim to meet current food needs but also ensure the preservation of resources for future generations.

This research project, titled "Global Rice Production Analysis," seeks to explore the foundational stages of data-driven research within the context of rice production. By leveraging extensive datasets, this study investigates key factors influencing rice yield and provides insights that can inform agricultural strategies.

Utilizing the dataset "Explore Data on Agricultural Production," which comprises approximately 13,500 rows and 40 columns, this research focuses on essential aspects of rice production and trade. The study addresses the following research questions:

How have global rice import and export trends evolved from 1980 to 2021, especially among the top five importing and exporting countries? Do importer and exporter countries differ in their rice production levels?

Which countries have produced over 25 million tons of rice annually in the past ten years, and which among them have improved the most? How do land use efficiency and production per capita compare between these high-producing countries and those that did not meet this threshold? Is there a significant difference in average rice production, land use efficiency, and production per capita between these groups?

What is the relationship between supply chain waste, domestic waste, and rice production in the top ten highest and ten lowest rice-producing countries? How can the

Waste to Production Ratio be utilized to assess efficiency in meeting domestic needs? Additionally, what is the extent of food waste occurring from production to consumer in these countries?

To ensure the reliability of the findings, the study begins with data cleaning, addressing issues such as missing values and outliers, followed by an examination of data distribution and interrelationships among key variables. By employing exploratory data analysis (EDA) and various statistical tests, this research not only addresses practical challenges faced by rice producers but also aims to enhance yield predictions.

Ultimately, the insights gained from this study will refine agricultural strategies, optimize resource allocation, and promote sustainable practices, contributing to the broader goal of strengthening global food security.

Dataset

1960-2022 for 190 countries and various regions

Body

Data Cleaning

The initial step in our analysis involved cleaning the dataset, which contained data for approximately 190 countries, along with continent-level data, European Union data, and some countries combined into groups by the FAO, all sourced from Our World in Data. We observed significant discrepancies between the mean and median values across all columns, primarily due to the aggregation of various countries into these groups. Such differences can lead to skewed results, prompting us to retain only the rows containing individual country data.

Next, we defined our time frame for analysis, focusing on the period from 1980 to

2020, which spans over four decades. This period included major rice crises, allowing us to investigate their impact on production trends.

To address missing data, we employed different imputation methods tailored to the type of variable. For population data, we calculated the average population growth rate for each country over the selected years and used this average to replace any missing values. This approach ensured that population estimates remained consistent with historical growth trends.

Exploratory Data Analysis(EDA)

We began by plotting histograms to examine the distribution of each column. The graphs revealed a right skew, prompting us to perform a log transformation to enhance symmetry. Following this, we plotted various columns against the production, our independent variable, to gain insights from the visualizations.

The production versus population scatterplot showed a general upward trend, indicating a positive correlation. However, the data points were widely dispersed, reflecting variability in how the population scales with higher production levels. While regions with higher production levels tend to have larger populations, notable exceptions exist. Most data points in the lower range suggest that most values fall within this segment. Similarly, a positive correlation was observed between production and land use, although this relationship was less strong and consistent. At higher production values, the increased scatter suggests that land use can vary considerably even for similar production levels. This variability may imply that factors beyond production influence land use, or that production efficiency—producing more with less land—differs across contexts.

Additionally, we examined correlations between production and other variables, reinforcing our earlier observations. The correlation matrix revealed notable relationships, with production showing a strong positive correlation with population and land use.

A QQ plot was generated for production, indicating that it approximately follows a normal distribution for most of the sample, with slight deviations observed in the tails.

Asnwering SMART question

To address the SMART questions, we first categorized countries into high and low producers based on their annual rice production. Countries producing more than 25 million tons of rice annually were classified as high producers. Out of the 190 countries in the dataset, only seven qualified as high producers: India, China, Indonesia, Bangladesh, Vietnam, Thailand, and Myanmar. Notably, all these countries are situated in Asia, indicating that factors such as soil type and climatic conditions may significantly influence their production capabilities. China leads with an average production of approximately 209 million tons, followed by India with around 166 million tons.

To determine which country has experienced the most improvement in annual production, we calculated the difference between production values at the end and the beginning of our observation period. Our analysis revealed that India has shown the most significant increase, with an enhancement of approximately 28.6 million tons over the specified timeframe.

Furthermore, we assessed land use efficiency, defined as the ratio of production to land use, alongside production per capita, which represents the ratio of production to the population of each country. These metrics will provide valuable insights into how effectively each country utilizes its land and resources in rice production.

Analysis of land Use Efficieny.   We plotted a box plot to examine differences in land use efficiency between high-producing and non-high-producing countries, revealing several key insights. The median land use efficiency for high producers is significantly higher than that of non-high producers, suggesting that countries consistently producing over 25 million tons of rice annually tend to utilize their land more efficiently. The spread

(interquartile range, or IQR) for high producers is narrower, indicating less variability in land use efficiency among these countries, which implies more uniform practices in land utilization for rice cultivation. In contrast, non-high producers exhibit a wider spread with several outliers, reflecting greater variability; while some may demonstrate high efficiency, the overall group lacks consistency. The presence of outliers within the non-high producers group indicates that a few countries achieve very high land use efficiency, but these instances are insufficiently consistent to qualify as high producers. To statistically validate the observed differences in land use efficiency, we conducted an independent samples t-test, which yielded a p-value of less than 0.05, indicating a statistically significant difference between the two groups

Analysis of Production per capita.   We plotted a box plot to examine differences in production per capita between high-producing and non-high-producing countries, revealing several key insights. The analysis of production per capita revealed several important findings. The median production per capita for high producers is significantly higher than that for non-high producers, suggesting that, on average, high-producing countries generate more rice per person. The spread (interquartile range, or IQR) for high producers is quite narrow, indicating less variability in production per capita and reflecting a more consistent level of production across these countries. In contrast, non-high producers exhibit a much wider spread with numerous outliers, indicating significant variability in production per capita; while most of these countries produce relatively low amounts per capita, some achieve notably higher figures. The presence of outliers in the non-high producers group suggests that certain smaller countries may produce substantial amounts of rice relative to their population size, likely due to intensive farming practices. To statistically validate these observed differences in production per capita, we conducted an independent samples t-test, which yielded a p-value of less than 0.05, indicating a statistically significant difference between the two groups.

Methods

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

Participants

Material

Procedure

Data analysis

We used R (Version 4.4.1; R Core Team, 2024) and the R-packages *papaja* (Version 0.1.2; Aust & Barth, 2023), and *tinylabels* (Version 0.2.4; Barth, 2023) for all our analyses.

Results

Discussion

References

Aust, F., & Barth, M. (2023). *papaja: Prepare reproducible APA journal articles with R Markdown.* Retrieved from https://github.com/crsh/papaja

Barth, M. (2023). *tinylabels: Lightweight variable labels.* Retrieved from https://cran.r-project.org/package=tinylabels

R Core Team. (2024). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/