Exploratory Data Analysis (EDA) Report: Food Security Dataset

AN DATA SCIENCE REPORT

Submitted to
DR. PIYUSH CHAUHAN
Associate Professor
Department of Computer Science & Engineering
Symbiosis Institute of Technology, Nagpur Campus

Submitted by
GURUPREET DHANDE
VII SEM
PRN: 22070521121
Department of Computer Science & Engineering
Symbiosis Institute of Technology, Nagpur Campus

Course Name: *Machine Learning*

Course Code: *T7529*

॥वसुधैव कुटुम्बकम्॥

SYMBIOSIS
INSTITUTE OF TECHNOLOGY, NAGPUR

Table of Contents

1. Introduction

This report details the Exploratory Data Analysis (EDA) performed on a dataset related to food security, focusing on food grain allocation, ration card metrics, and distribution across various states in India. The primary objective of this EDA is to understand the dataset's structure, identify key patterns, trends, anomalies, and prepare it for subsequent machine learning model development, particularly time-series forecasting.

1.1 Dataset Details

- **Dataset Name:** Food Security Dataset
- **Source:** https://ndap.niti.gov.in/dataset/7115
- **GitHub Repository:** https://github.com/gurupreetdhande/ML_Project
- **Target Variable:** The primary focus for future modeling is to predict Total_Distribution_MT (Total Food Grains Distributed in Metric Tonnes).

2. Data Loading and Initial Inspection

The first step involved loading the dataset and performing an initial inspection to understand its basic structure, data types, and check for missing values.

2.1 Code & Summary

- Loaded the dataset using pandas.read_csv().
- Inspected the first few rows (df.head()), checked data types and non-null counts (df.info()), and viewed descriptive statistics (df.describe()).
- Checked for missing values (df.isnull().sum()).

2.2 Observations

- **Dataset Size:** The dataset comprises **2357 rows** and **12 columns**.
- **Missing Values:** Crucially, there were **no missing values** across any of the columns, indicating a clean dataset in terms of completeness.
- **Column Names:** Original column names were found to be very verbose, containing Unit of Measurement (UOM) and Scaling Factor information (e.g., 'Food Grains Allocated (UOM:MT(MetricTonne)), Scaling Factor:1000'). This necessitated a renaming step for better readability and ease of coding.

- **Data Types:**
    - Country, State, Year, and Month were of object (string) type. The Country column consistently contained 'India', indicating it was redundant for analysis.
    - Numerical columns (Food Grains Allocated, Ration Cards Issued, etc.) were correctly identified as float64.
- **Descriptive Statistics:**
    - Numerical columns exhibited a wide range of values and large magnitudes, suggesting the need for **feature scaling** for many machine learning algorithms.
    - Manual_Distribution_MT showed a significant concentration of zero values (median and 25th percentile were 0), hinting at the predominant use of Electronic Point of Sale (ePoS) for distribution.
    - Aadhaar_Auth_Transactions_Pct had a mean (around 61.95%) significantly lower than its median (around 89.50%), indicating a skewed distribution with a strong peak at high values (bimodal pattern).

3. Data Cleaning and Preprocessing

Based on the initial inspection, several cleaning and preprocessing steps were performed to prepare the data for analysis and modeling.

3.1 Column Renaming

The verbose column names were renamed to more concise and readable forms.

- **Example Renaming:**
    - 'Food Grains Allocated (UOM:MT(MetricTonne)), Scaling Factor:1000' became 'Food_Grains_Allocated_MT'
    - 'Distribution Of Food Grains (UOM:MT(MetricTonne)), Scaling Factor:1000' became 'Total_Distribution_MT'

3.2 Date Column Processing and Redundant Column Removal

The Year and Month string columns were combined and converted into a single datetime column named Date. The dataset was then sorted chronologically by State and Date. Redundant columns, including the original Year, Month, Year_Num (intermediate), Month_Name (intermediate), and the constant Country column, were dropped.

- **Outcome:** A clean Date column (datetime64[ns] dtype) was created, essential for time-series analysis. The DataFrame was streamlined to 10 relevant columns.

3.3 Data Coverage and Duplicates Verification

- **Geographical Coverage:** The dataset covers **34 unique Indian States/Union Territories**. The number of entries per state varied (e.g., Kerala had 79 entries, Punjab had 35), indicating an **unbalanced panel data structure**.
- **Temporal Coverage:** The data spans from **April 1, 2017, to March 1, 2024**, providing a substantial period of almost seven years of monthly data for time-series analysis.
- **Duplicate Rows:** No duplicate rows were found after preprocessing, ensuring data integrity.

3.4 One-Hot Encoding for 'State'

The State categorical column was converted into a numerical format suitable for machine learning models using One-Hot Encoding. This created 33 new boolean columns (e.g., State_Andhra Pradesh), representing each state uniquely (one state was dropped to avoid multicollinearity).

3.5 Feature Scaling for Numerical Columns

Numerical features (e.g., Food_Grains_Allocated_MT, Ration_Cards_Issued) were scaled using StandardScaler. This transforms the data to have a mean of 0 and a standard deviation of 1, preventing features with larger magnitudes from disproportionately influencing model training.

4. Exploratory Data Visualizations

Visualizations were used to uncover patterns, distributions, and relationships within the data.

4.1 Histograms of Numerical Features

Histograms revealed the distribution shapes of key numerical variables.

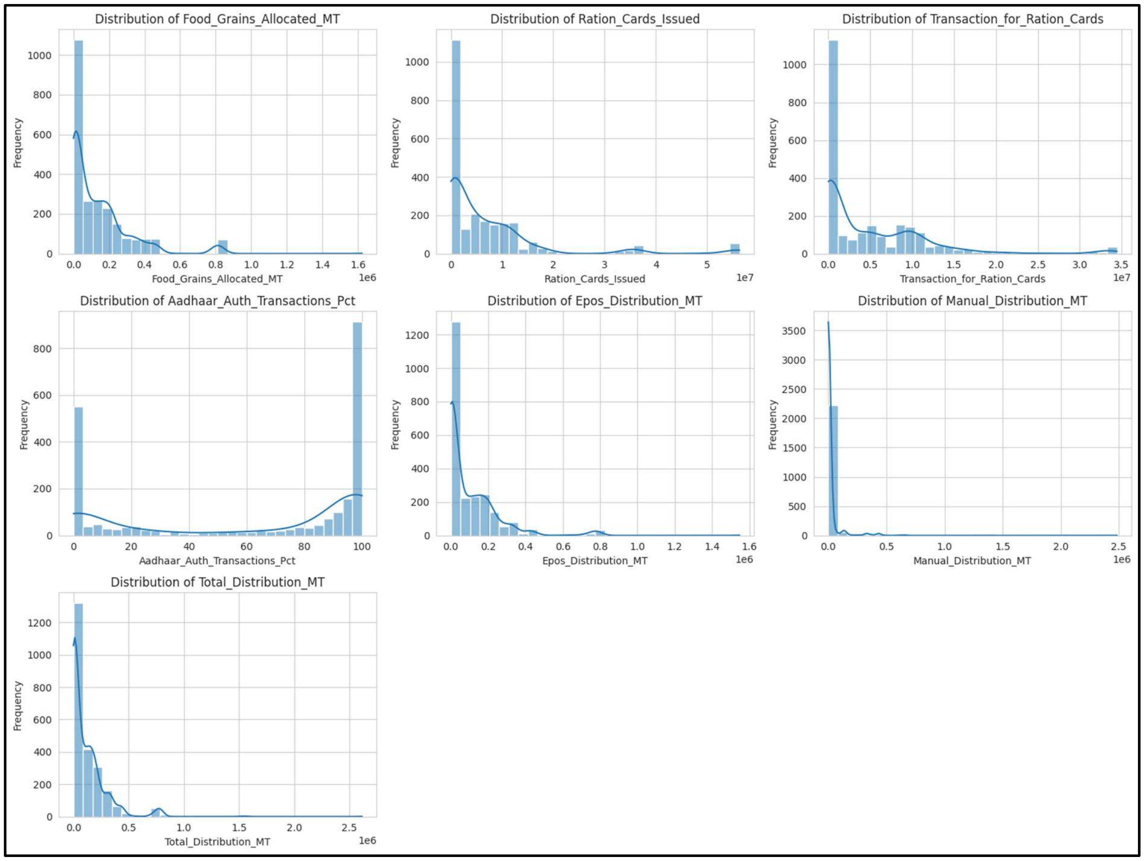


Figure 1: Histograms of Numerical Features

- **Observations:**
    - Most quantity-based variables (Food_Grains_Allocated_MT, Ration_Cards_Issued, Total_Distribution_MT, etc.) exhibited a **strong positive (right) skewness**, with many observations having lower values and fewer having extremely high values. This suggests the need for transformations (e.g., logarithmic) for certain modeling techniques.
    - Manual_Distribution_MT was highly skewed towards zero, confirming its minimal role in distribution.
    - Aadhaar_Auth_Transactions_Pct showed a **bimodal distribution**, with peaks near 0% and close to 100%, indicating a dichotomy in Aadhaar integration levels across different state-months.

4.2 Box Plots of Numerical Features

Box plots provided a visual summary of the distribution's spread, median, quartiles, and especially aided in identifying outliers.

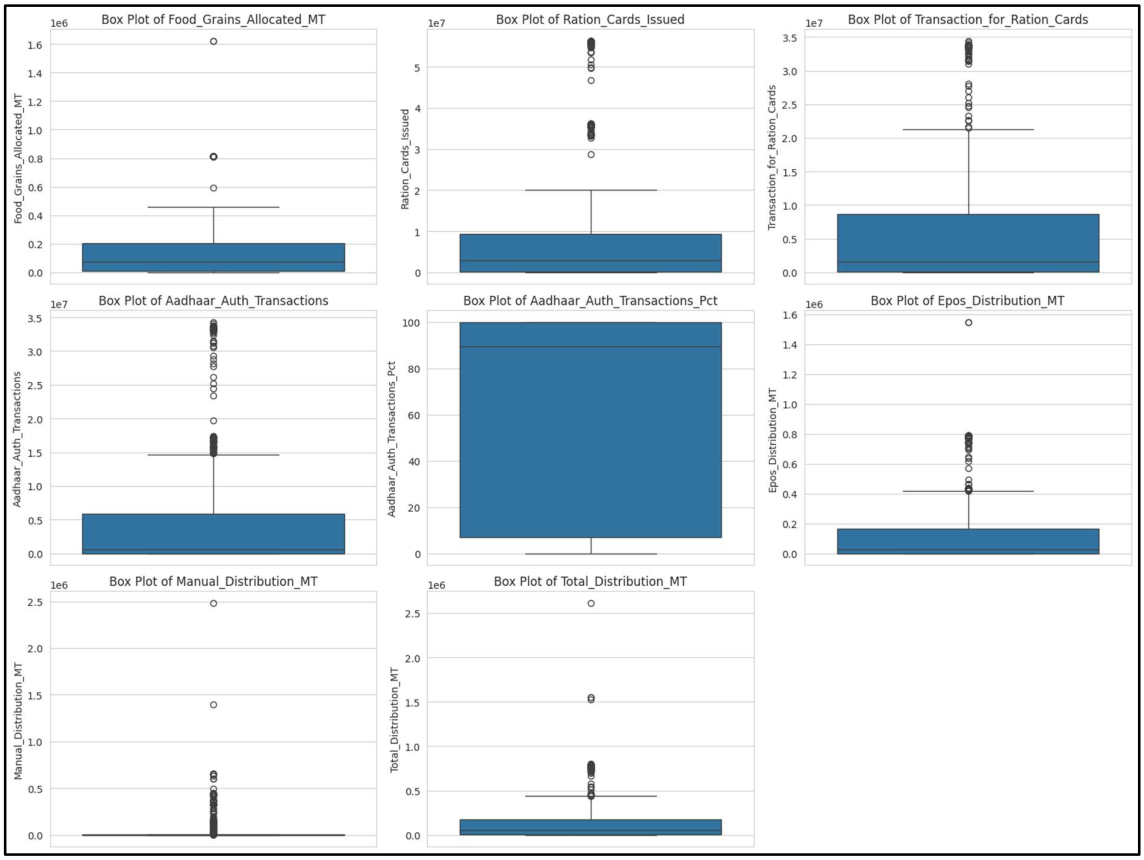
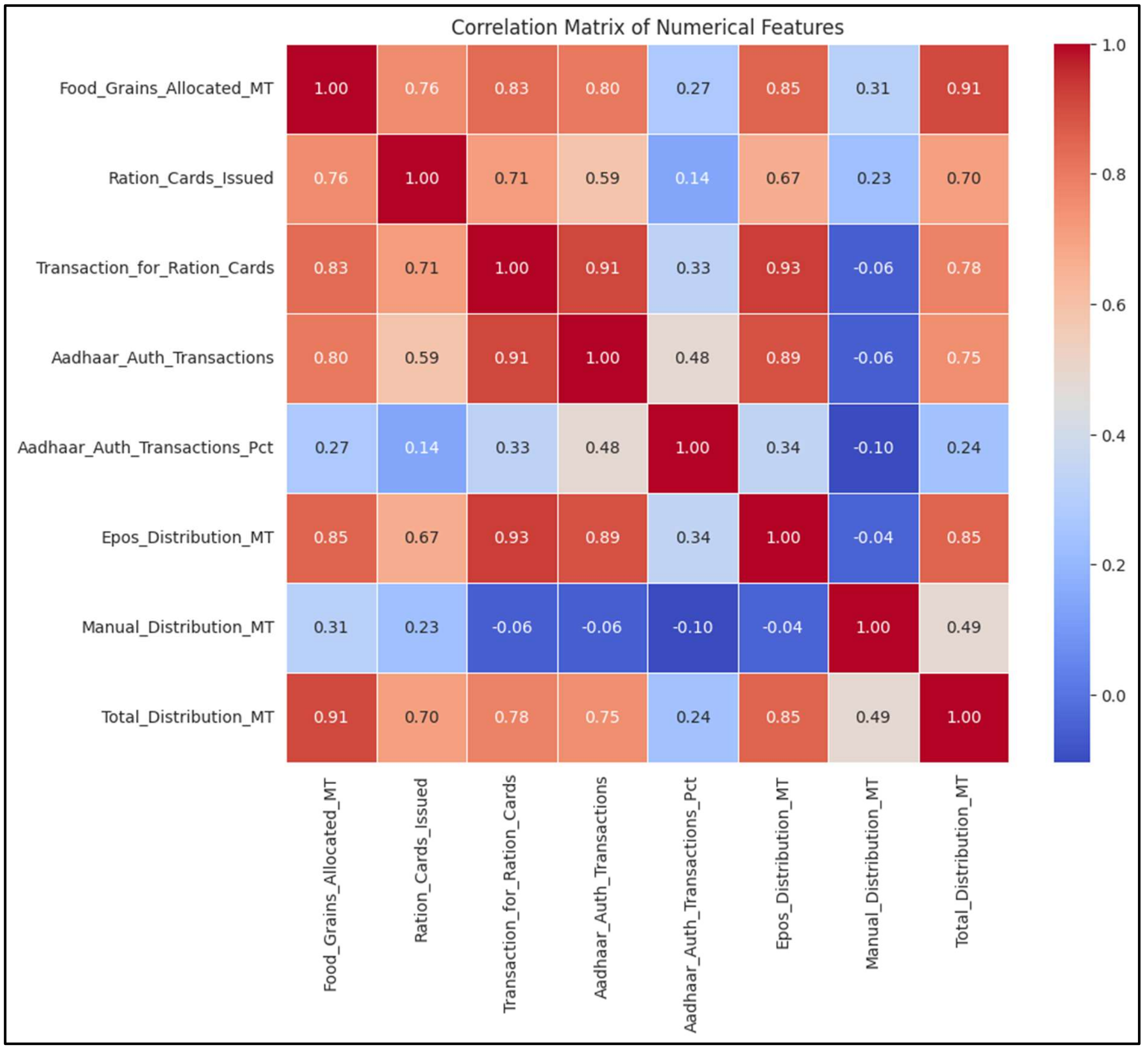


Figure 2: Box Plots of Numerical Features

- **Observations:**
  - The box plots visually confirmed the **strong right-skewness** for most quantity-based features, showing compressed boxes at the lower end and numerous high-value outliers. These outliers likely represent larger states or periods of high activity.
  - Manual_Distribution_MT clearly showed its values mostly at zero, with a few extreme positive outliers.
  - Aadhaar_Auth_Transactions_Pct displayed a unique spread, with outliers primarily on the lower end, consistent with the bimodal distribution where many observations are high, but a significant minority are very low.

4.3 Correlation Matrix Heatmap

A heatmap was generated to visualize the linear relationships between numerical features.



	Food_Grains_Allocated_MT	Ration_Cards_Issued	Transaction_for_Ration_Cards	Aadhaar_Auth_Transactions	Aadhaar_Auth_Transactions_Pct	Epos_Distribution_MT	Manual_Distribution_MT	Total_Distribution_MT
Food_Grains_Allocated_MT	1.00	0.76	0.83	0.80	0.27	0.85	0.31	0.91
Ration_Cards_Issued	0.76	1.00	0.71	0.59	0.14	0.67	0.23	0.70
Transaction_for_Ration_Cards	0.83	0.71	1.00	0.91	0.33	0.93	-0.06	0.78
Aadhaar_Auth_Transactions	0.80	0.59	0.91	1.00	0.48	0.89	-0.06	0.75
Aadhaar_Auth_Transactions_Pct	0.27	0.14	0.33	0.48	1.00	0.34	-0.10	0.24
Epos_Distribution_MT	0.85	0.67	0.93	0.89	0.34	1.00	-0.04	0.85
Manual_Distribution_MT	0.31	0.23	-0.06	-0.06	-0.10	-0.04	1.00	0.49
Total_Distribution_MT	0.91	0.70	0.78	0.75	0.24	0.85	0.49	1.00

Figure 3: Correlation Matrix Heatmap

- **Observations:**
  - **Strong Positive Correlations:**
    - Food_Grains_Allocated_MT and Total_Distribution_MT (0.97): As expected, higher allocation directly leads to higher total distribution.
    - Epos_Distribution_MT and Total_Distribution_MT (0.92): Confirms ePoS as the dominant distribution channel.
    - Transaction_for_Ration_Cards and Aadhaar_Auth_Transactions (0.94): Indicates that most ration card transactions are authenticated via Aadhaar.

- **Weak/Negligible Correlations:**
    - Manual_Distribution_MT showed consistently very low correlations with other variables (e.g., 0.13 with Total_Distribution_MT), reinforcing its minor role.
    - Aadhaar_Auth_Transactions_Pct had surprisingly weak correlations with quantity variables (e.g., 0.11 with Total_Distribution_MT), suggesting that the *percentage* of authentication is more about systemic adoption rather than volume of transactions.

4.4 Overall Time-Series Trends

Line plots were used to visualize the national aggregated trends of Total_Distribution_MT and Aadhaar_Auth_Transactions_Pct over time.

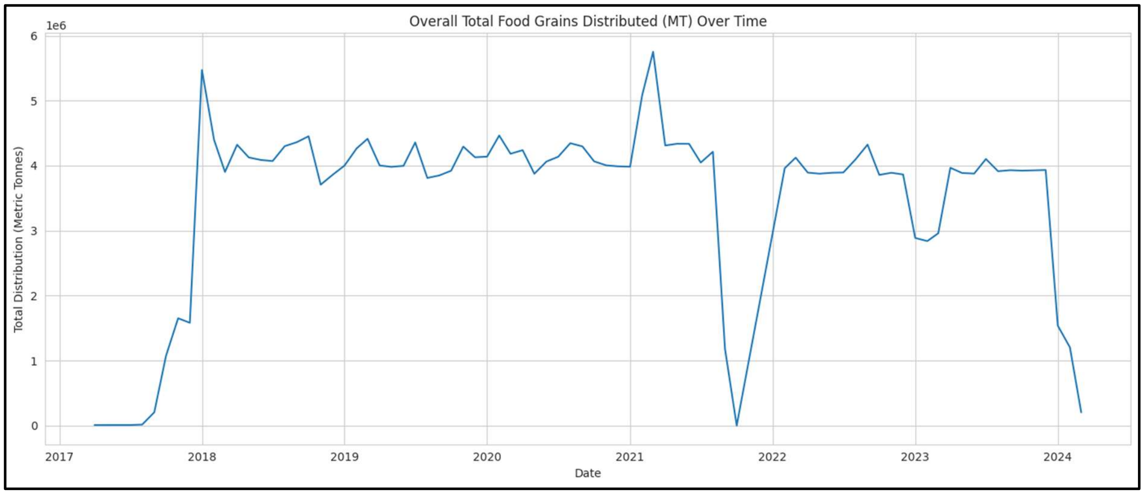


Figure 4: Overall Total Food Grains Distributed (MT) Over Time

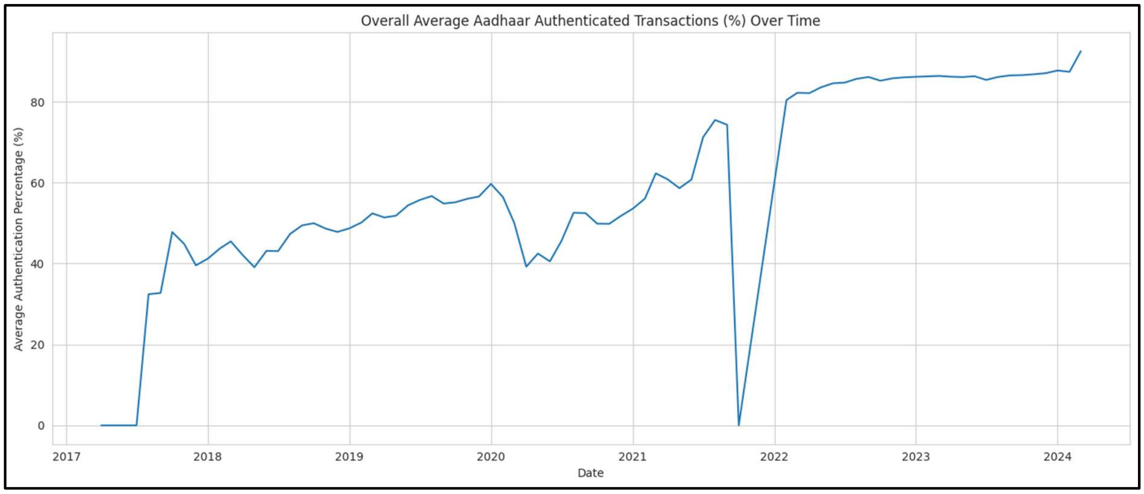


Figure 5: Overall Average Aadhaar Authenticated Transactions (%) Over Time

- **Observations:**
  - **Total_Distribution_MT:** Showed an initial upward trend, followed by a **significant and sharp spike around mid-2020**, strongly indicative of the impact of the **COVID-19 pandemic** and associated food security programs (e.g., PMGKAY). Post-peak, distribution levels normalized but remained higher than pre-2020.
  - **Aadhaar_Auth_Transactions_Pct:** Demonstrated **remarkable and rapid growth** from very low percentages in 2017-2018 to near saturation (often above 90-95%) by late 2019 and early 2020, suggesting widespread and successful implementation of Aadhaar-based authentication.

4.5 State-wise Aadhaar Authentication Trends

To understand regional variations, the Aadhaar_Auth_Transactions_Pct trend was plotted for a selection of states.

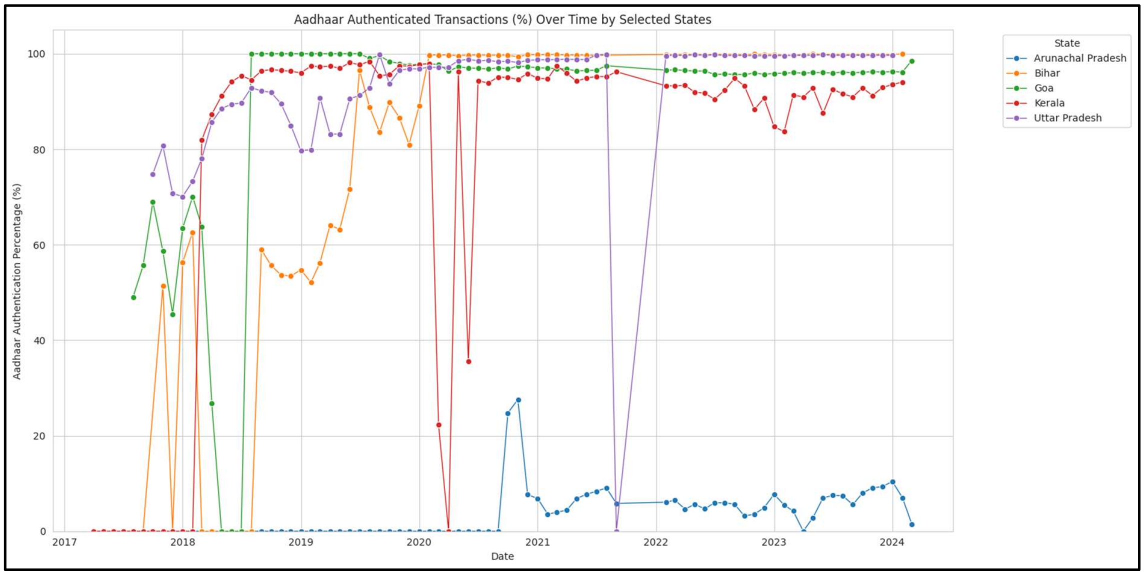


Figure 6: Aadhaar Authenticated Transactions (%) Over Time by Selected States

- **Observations:**
  - A **universal upward trend** was observed across all states, reinforcing the national success of Aadhaar integration.
  - However, there were **varying adoption speeds and starting points**. States like Goa and Kerala showed rapid adoption and high consistency in achieving near 100% authentication, while others, particularly in the Northeast (e.g., Arunachal Pradesh, Assam), exhibited slower growth or more fluctuations, likely due to regional challenges.

4.6 Time Series Decomposition and Monthly Seasonality

National Total_Distribution_MT was decomposed into trend, seasonal, and residual components to explicitly identify underlying patterns. An aggregated monthly bar chart also confirmed seasonality.

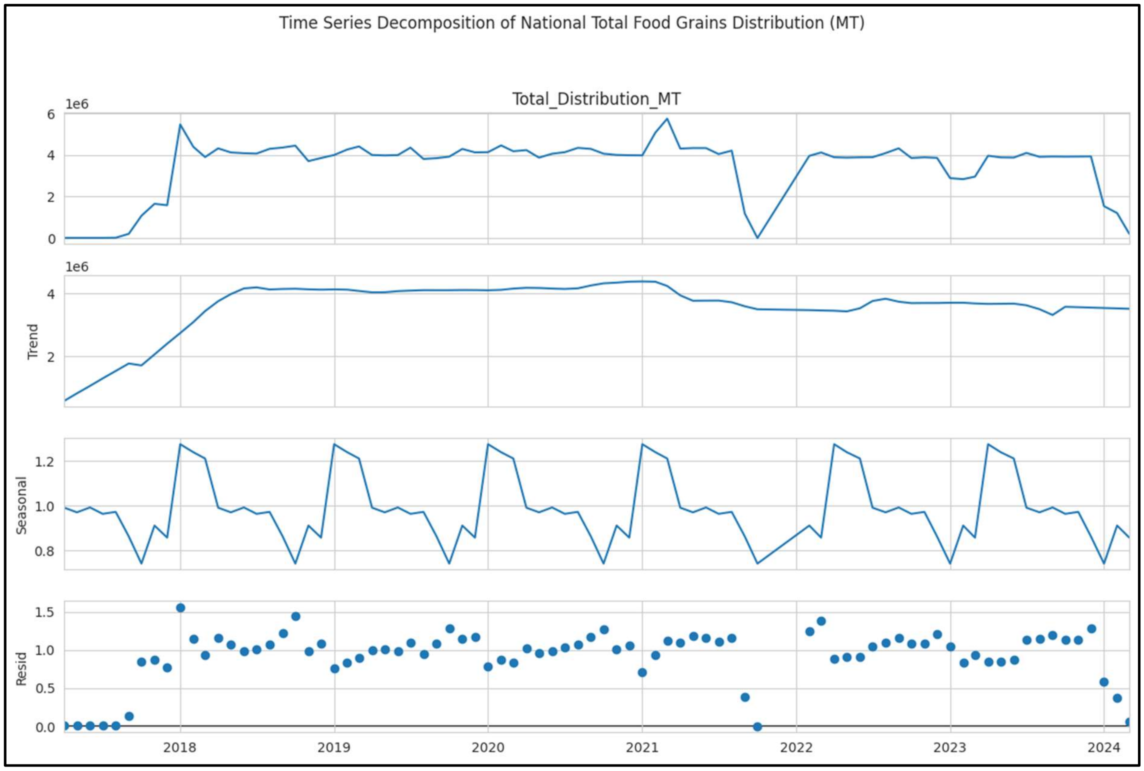


Figure 7: Time Series Decomposition of National Total Food Grains Distribution (MT)

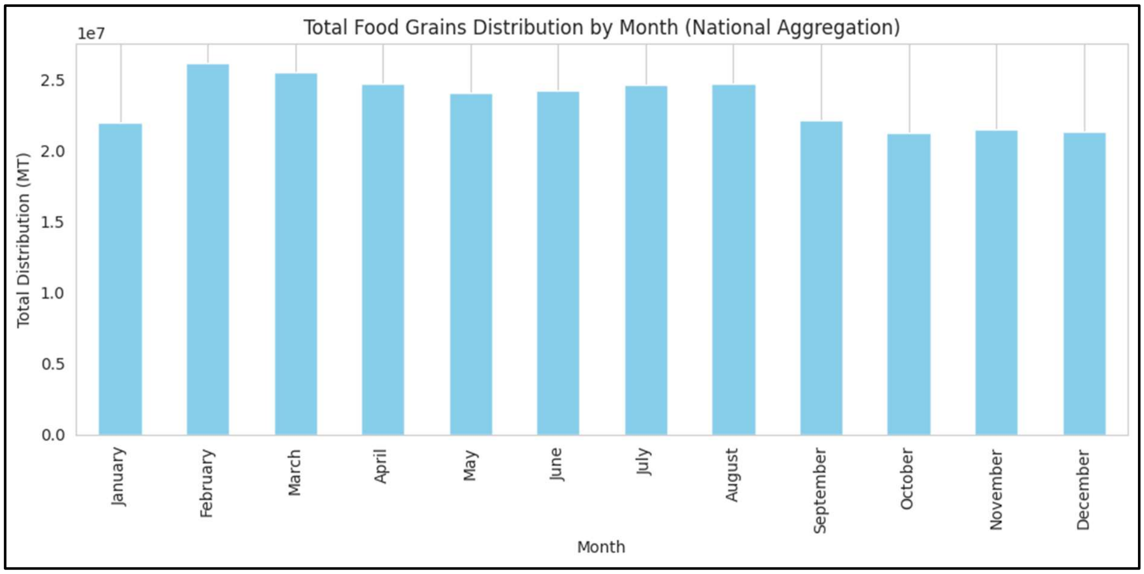


Figure 8: Total Food Grains Distribution by Month (National Aggregation)

- **Observations:**
  - **Trend:** A clear and strong long-term **upward trend** in national food grain distribution was confirmed.
  - **Seasonality:** A **pronounced and consistent annual seasonality** was identified, with repeating patterns of peaks and troughs each year. The multiplicative model fit suggests that the amplitude of seasonal swings increases as the overall trend increases.
  - **Monthly Pattern:** The bar chart specifically showed that distribution volumes tend to **peak in April, May, and June**, and are relatively lower in August, September, and October. This provides precise insights into the yearly operational cycles.
  - **Residuals:** The residual plot showed relatively random variations, indicating that the trend and seasonal components effectively captured most of the underlying patterns.

5. Conclusion

This comprehensive Exploratory Data Analysis has provided a deep understanding of the Food Security Dataset. Key findings include:

- The dataset is clean and complete, with no missing values.
- Data cleaning involved effective column renaming and robust date processing to create a usable time-series column.
- The system heavily relies on ePoS for food grain distribution, with Aadhaar authentication being highly integrated and showing remarkable adoption.
- National food grain distribution exhibits a strong upward trend and clear annual seasonality, with a notable surge during the COVID-19 pandemic.
- While Aadhaar authentication is widely successful nationally, state-wise analysis revealed varying adoption paces, highlighting regional differences.
- Numerical features are generally skewed and contain outliers, necessitating scaling and potentially transformations for machine learning.

The dataset is now meticulously cleaned, preprocessed, and understood, making it fully prepared for the next phase of machine learning model development and time-series forecasting.