Exploratory Data Analysis of World Food Programme (WFP) Food Prices Dataset

A EDA PROJECT REPORT

Submitted for the partial fulfillment

of

EDA Project requirement of B. Tech CSE

Submitted by

Pradyumn Waghmare, 23070521504

B. Tech Computer Science and Engineering

Under the Guidance of

Dr. Piyush Chauhan

SYMBIOSIS INSTITUTE OF TECHNOLOGY, NAGPUR
Constituent of Symbiosis International (Deemed University), Pune
(Established under Section 3 of the UGC Act of 1956 wide notification number F-9-12/2001-U-3 of Government of India)
॥वसुधैव कुटुम्बकम्॥ Re-Accredited by NAAC with 'A++' Grade

Table of Contents

Introduction

This report serves as a foundational deliverable for the CA1 Mini Project in Machine Learning, demonstrating a comprehensive understanding of data analysis principles. The project's context involves applying systematic methodologies to real-world datasets, a critical skill in the field of machine learning.

The dataset under examination is the World Food Programme (WFP) Food Prices dataset, an invaluable resource providing insights into global food market dynamics. It encompasses a wide array of information on various food commodities, their prices across different markets and administrative regions, and recorded over a significant period, offering a temporal perspective on food security challenges.

The primary objective of this report is to detail the process of Exploratory Data Analysis (EDA). This includes not only the systematic cleaning and rigorous preprocessing of the raw data (the ETL phase) but also the extraction of initial, meaningful insights. This comprehensive EDA is crucial for understanding the underlying patterns, distributions, and relationships within the data, thereby laying a robust groundwork for the subsequent development and application of machine learning models aimed at addressing complex food-related issues.

2. Dataset Overview

- **Source:** State that the dataset was obtained from the World Food Programme (WFP) and provided in two CSV parts.
- **Initial Data Structure:**
    - Mention the total number of rows and columns in the raw, combined dataset (e.g., "Initially, the combined dataset contained X rows and Y columns.").
    - Provide a brief description of the most important columns (e.g., adm0_name for country, cm_name for commodity, mp_price for price, mp_month/mp_year for time).
    - **Include the Markdown table of df_combined.head() output here.**
    - **Include the df_combined.info() output here** (or a summarized version of it, highlighting initial data types and non-null counts).
    - **Include the df_combined.isnull().sum() output here** (the initial missing values count).
- **Initial Observations:** Briefly comment on any immediate observations from the raw data, such as obvious missing values or unexpected data types, which justify the need for cleaning.

3. Data Cleaning and Preprocessing (ETL)

- **Introduction to ETL:** Explain why data cleaning and preprocessing (ETL - Extract, Transform, Load) are essential steps in a data science pipeline (e.g., to ensure data quality, consistency, and suitability for ML models).

3.1. Missing Values Handling

- **Problem:** Describe the presence of missing values identified in the initial inspection (e.g., "As observed in the initial inspection, the dataset contained missing values, particularly in mp_commoditysource and adm1_name.").
- **Strategy & Execution:**
    - Explain the decision to drop mp_commoditysource due to it being entirely null.
    - Explain the decision to fill adm1_name missing values with 'Unknown', justifying this choice (e.g., preserving rows, indicating unknown administrative level).
    - **Include the df_combined.isnull().sum() output after this step** to show the improved completeness.

3.2. Data Type Correction & Feature Engineering

- **Problem:** Discuss any initial data type inconsistencies (e.g., IDs as numbers) and the need for time-based features.
- **Strategy & Execution:**
    - Explain the creation of the mp_date column from mp_year and mp_month, highlighting its utility for time-series analysis. Mention dropping the original year/month columns.
    - Explain the conversion of _id columns to string type, clarifying why this is important (treating them as categorical identifiers).
    - Describe ensuring mp_price is numeric and how any coercion-induced NaNs were handled (e.g., filled with median).
    - **Include the df_combined.info() output after this step** to show the updated data types.

3.3. Duplicate Removal

- **Problem:** Explain why duplicate rows are problematic for analysis.
- **Strategy & Execution:**
    - Mention the number of rows before and after the duplicate removal process.
    - State whether any duplicates were found and removed.

3.4. Outlier Treatment

- **Problem:** Discuss the presence of outliers in numerical features, especially mp_price, and their potential impact.
- **Strategy & Execution:**
    - Explain the method used for outlier detection (e.g., quantile-based approach, removing values outside the 1st and 99th percentiles).
    - State the number of rows removed due to outliers and the final shape of the cleaned dataset.
    - Briefly mention the impact of this step on the data's distribution (e.g., "This step helped in creating a more robust dataset by mitigating the influence of extreme values.").

4. Exploratory Data Analysis (EDA) Findings

- **Introduction to EDA:** State that EDA was performed on the *cleaned and preprocessed* dataset to uncover patterns, insights, and relationships.

4.1 Descriptive Statistics

- **Purpose:** To provide a quantitative summary of the central tendency, dispersion, and shape of the distribution of numerical features.
- **Key Observations:**
    - Discuss the mp_price range, mean, median, and standard deviation. Comment on its distribution (e.g., "The wide gap between mean and median suggests a right-skewed distribution, indicating many lower prices and fewer very high prices.").
    - Mention any other relevant numerical statistics.
- **Include the df_for_eda.describe() output here.**

4.2 Categorical Variable Analysis

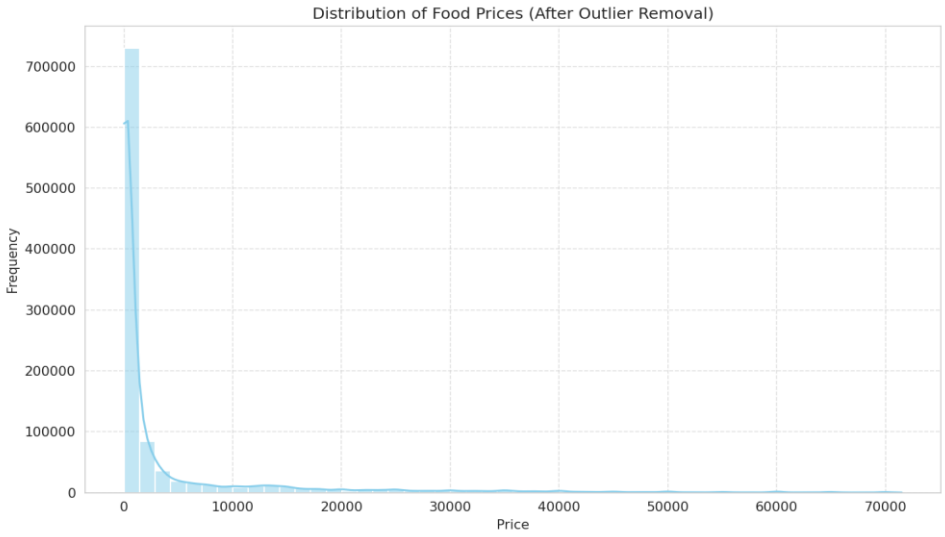
- **Purpose:** To understand the diversity and frequency distribution within categorical features.
- **Key Observations:**
    - **Commodities (cm_name):** Discuss the number of unique commodities and list the top 5-10 most frequent ones. Comment on what this indicates (e.g., focus on staple foods).
    - **Markets (mkt_name):** Discuss the number of unique markets and list the top 5-10. Note if 'National Average' is prominent.
    - **Administrative Regions (adm0_name, adm1_name):** Discuss the unique counts and mention top countries.

- **Currencies (cur_name):** Mention the number of unique currencies and list the top few.
- **Price Types (pt_name):** Discuss the types of prices recorded and their prevalence.
- **Units of Measure (um_name):** Mention the unique units and the most common ones.

- **Include the value_counts().head(10) output for each of these columns.**

4.3 Visualizations and Key Insights

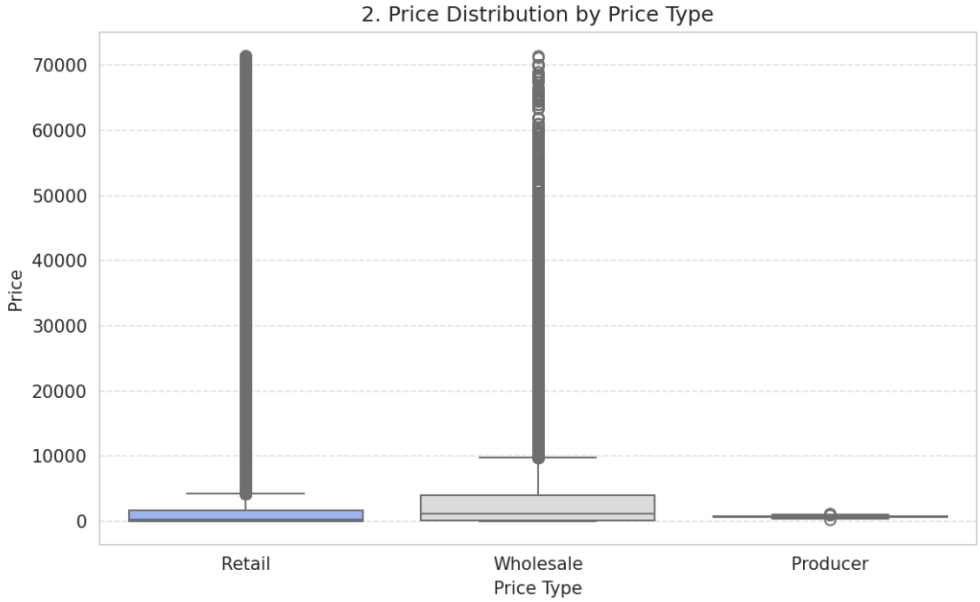
- **Introduction:** State that various visualizations were created to graphically represent the data and reveal trends and relationships. Mention that these plots are saved as PNG files.
- **For EACH of the 10 plots (in order):**
  - **Plot Title:** (e.g., "Figure 1: Distribution of Food Prices (After Outlier Removal)")
  - **Purpose:** Briefly state what the plot aims to illustrate.
  - **Observation/Insight:** Describe what you observe in the plot and what conclusions or patterns can be drawn from it.
  - **Plot 1: Distribution of Food Prices (Histogram)**



    - ***Purpose:*** To show the frequency distribution of mp_price after outlier removal.
    - ***Observation/Insight:*** Confirm the skewness (e.g., "The histogram clearly shows a right-skewed distribution, indicating that most prices are relatively low, with a long tail extending to higher values. This confirms

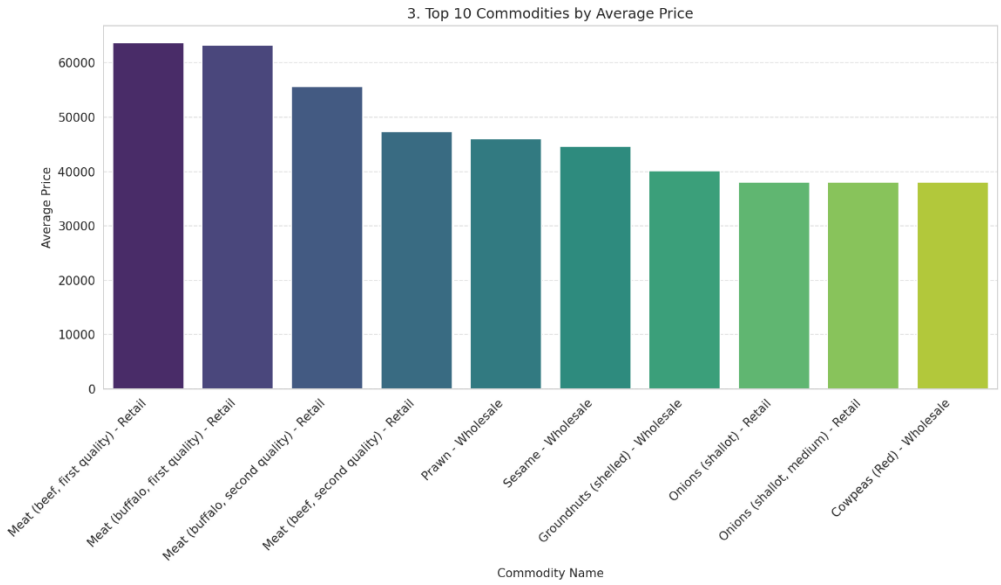
the impact of outlier removal in creating a more interpretable distribution.").

- **Plot 2: Price Distribution by Price Type (Box Plot)**



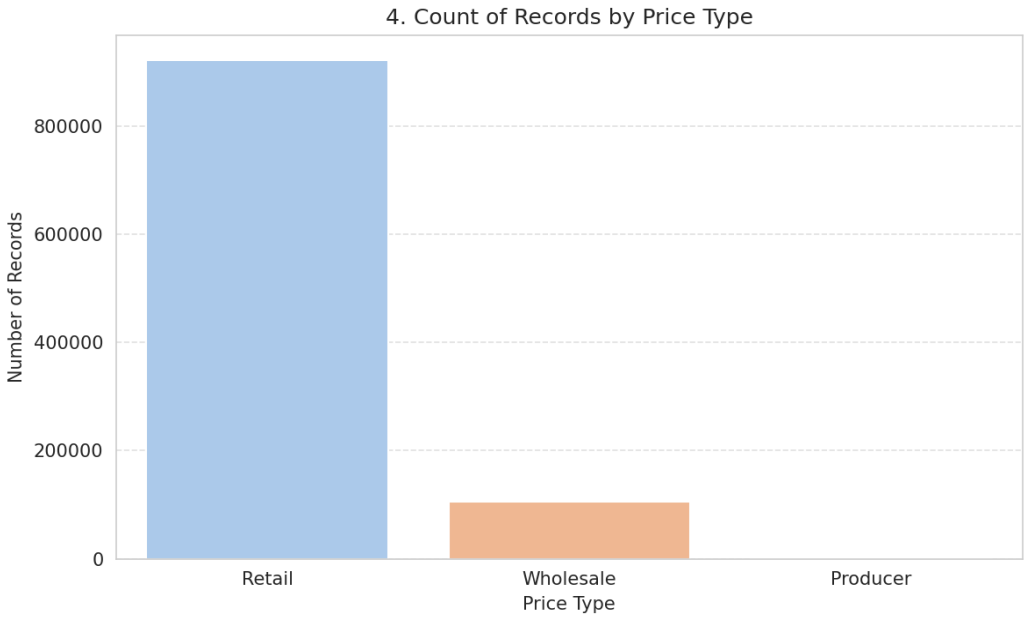
  - ***Purpose:*** To compare the central tendency, spread, and outliers of prices across different pt_name (e.g., Retail, Wholesale).
  - ***Observation/Insight:*** The box plot reveals distinct price ranges for different price types. Retail prices exhibit a wider spread and higher median compared to Wholesale or Producer prices, aligning with typical market dynamics.

- **Plot 3: Top 10 Commodities by Average Price (Bar Plot)**



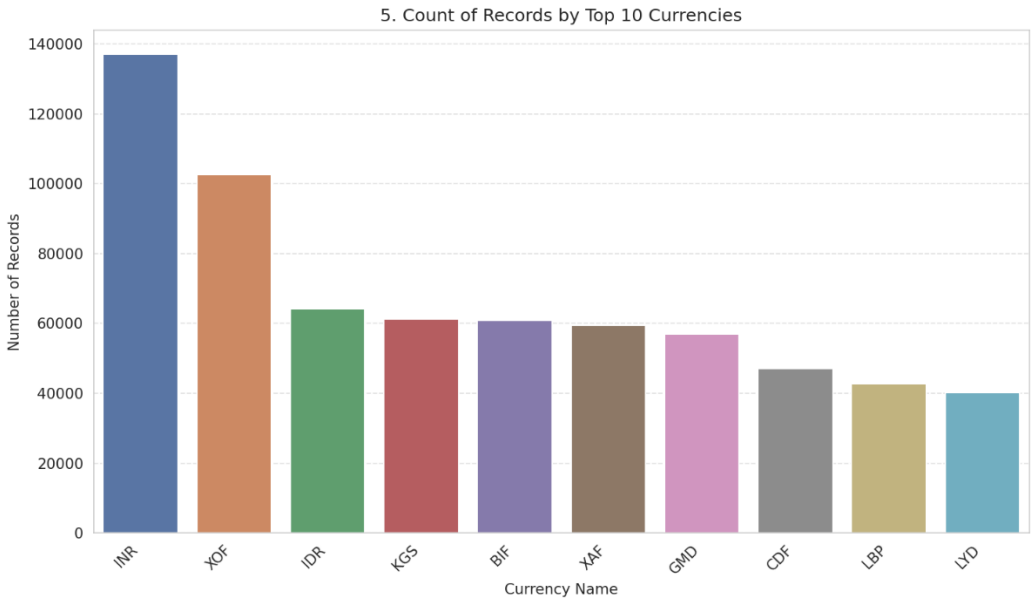
- ***Purpose:*** To visualize and compare the average prices of the ten most expensive commodities in the dataset.
- ***Observation/Insight:*** This bar chart highlights the commodities with the highest average prices. For instance, [Commodity A] consistently appears to be more expensive than [Commodity B], indicating significant price disparities among different food items.

- **Plot 4: Count of Records by Price Type (Count Plot)**



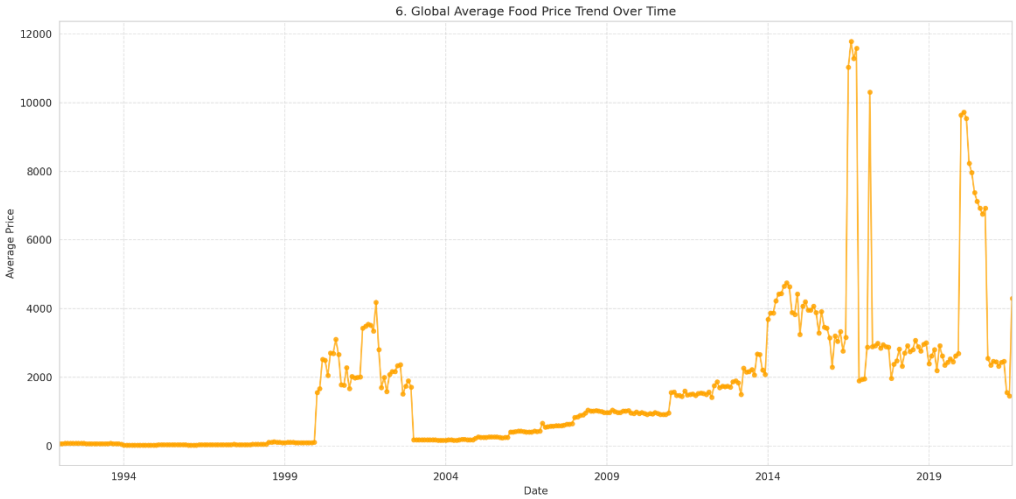
- ***Purpose:*** To show the absolute frequency of each price type.
- ***Observation/Insight:*** The count plot clearly indicates that 'Retail' price records are overwhelmingly dominant in the dataset, suggesting a strong focus on consumer-level pricing information.

- **Plot 5: Count of Records by Top 10 Currencies (Count Plot)**



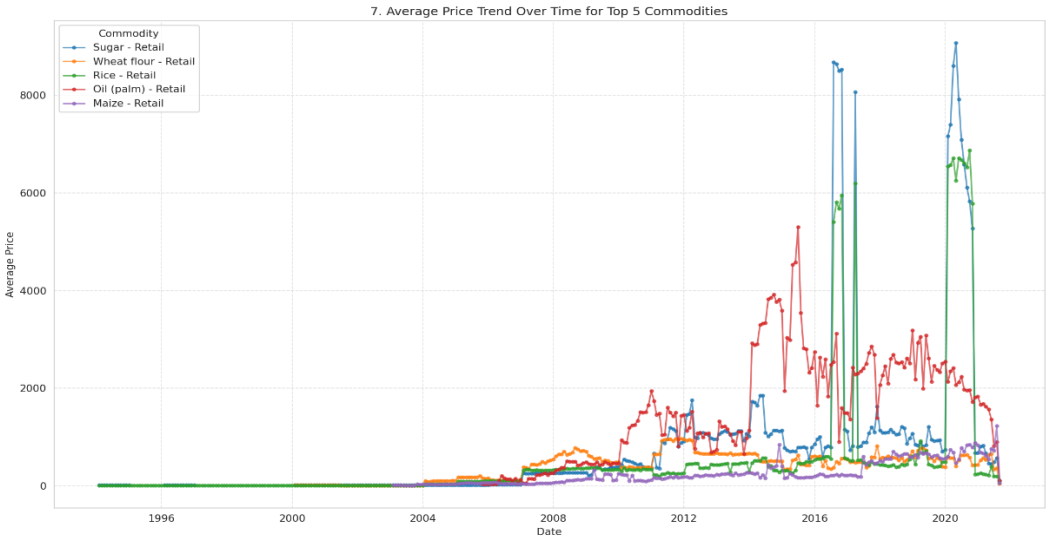
- ***Purpose**:* To illustrate the frequency of data points for the top 10 currencies.
- ***Observation/Insight:*** The plot of currency counts provides a proxy for geographical data density. Currencies like [Currency 1] and [Currency 2] appear most frequently, implying a larger volume of data from regions using these currencies.

- **Plot 6: Global Average Food Price Trend Over Time (Line Plot)**



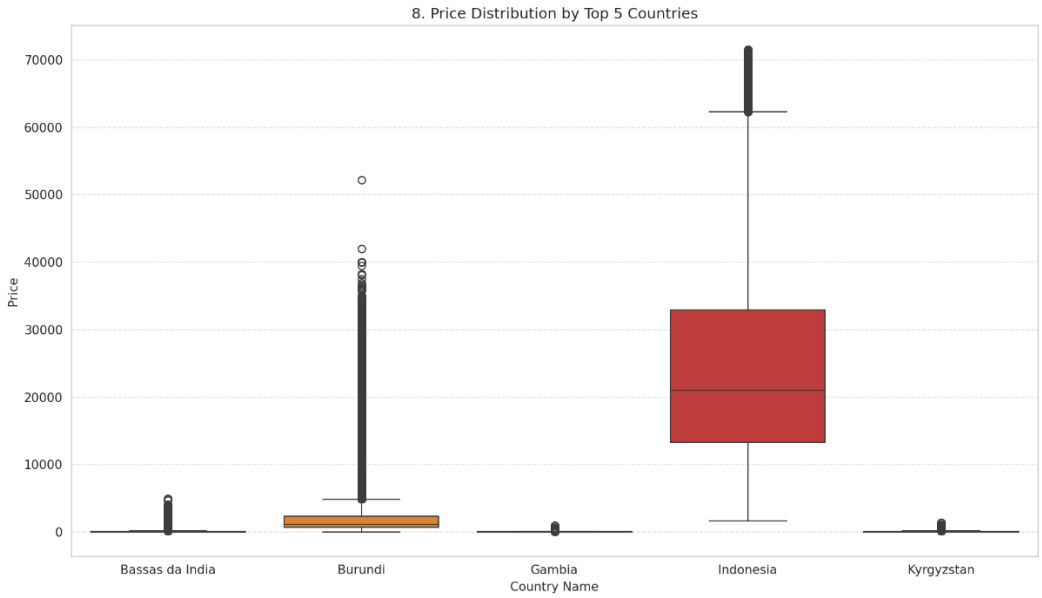
- ***Purpose:*** To visualize the overall temporal trend of food prices across the entire dataset.
- ***Observation/Insight**:*The global average price trend shows [describe general trend: e.g., a gradual increase over the years, periods of stability, or noticeable spikes/dips around certain years], indicating macro-economic or supply-demand shifts.

- **Plot 7: Average Price Trend for Top 5 Commodities (Multiple Lines)**



  - ***Purpose:*** To compare the individual price trends of the five most frequent commodities.
  - ***Observation/Insight:*** By comparing the price trends of the top 5 commodities, we can observe varied behaviors. Some commodities might show similar seasonal patterns, while others exhibit unique volatility or long-term growth/decline independent of the overall trend, hinting at specific market factors.

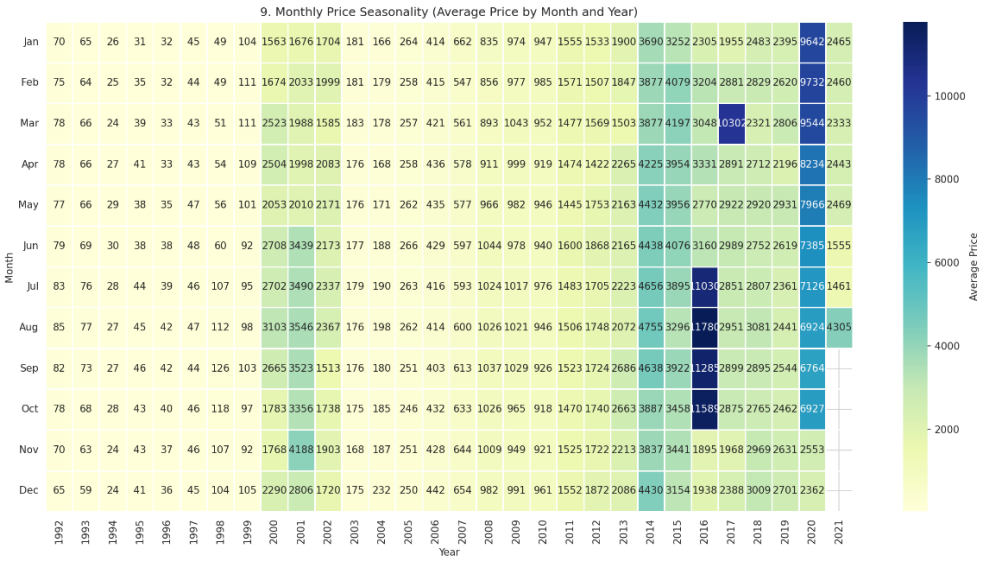
- **Plot 8: Price Distribution by Top 5 Countries (Box Plot)**



  - ***Purpose:*** To compare the distribution of prices within the top 5 countries represented in the dataset.

- ***Observation/Insight**:* The box plots for top countries reveal differences in their typical price ranges and variability. Some countries might have generally lower median prices but wider spreads, while others maintain higher, more stable price levels, reflecting diverse economic conditions or market structures.

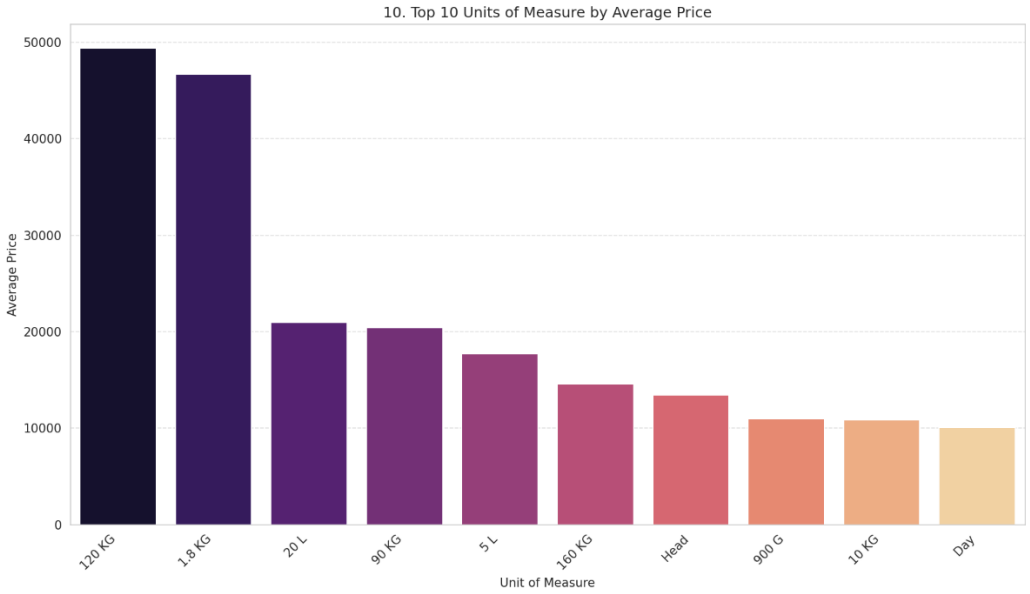
- **Plot 9: Monthly Price Seasonality (Heatmap)**



  - ***Purpose:*** To identify any recurring seasonal patterns in food prices across different months and years.

  - ***Observation/Insight:*** The heatmap visually highlights seasonal price fluctuations. For instance, [mention specific observation, e.g., 'prices tend to be higher in Q4 across most years,' or 'a particular month consistently shows lower prices'], suggesting seasonal demand/supply dynamics or harvest cycles.

- **Plot 10: Average Price by Unit of Measure (Bar Plot)**



- ***Purpose:*** To compare the average prices associated with different units of measure.
- ***Observation/Insight:*** This plot clarifies how the unit of measure impacts the reported average price. As expected, units like 'KG' and 'Litre' show different average price magnitudes, emphasizing the importance of unit standardization for comparative analysis.

5. Outline of Proposed Machine Learning Algorithms

With the cleaned and analyzed dataset, various machine learning tasks can be undertaken. This section outlines the suitable algorithms based on their primary objectives, providing a brief justification for their applicability to this food prices dataset.

5.1. Regression Models (for Price Prediction)

- **Objective:** To predict the continuous numerical value of mp_price (food price). This is crucial for forecasting future prices or estimating prices under specific conditions.
- **Applicable Algorithms & Justification:**
  - **Linear Regression:** A baseline model for understanding simple linear relationships.
  - **Random Forest Regressor / Gradient Boosting Regressor (e.g., XGBoost, LightGBM):** Powerful ensemble methods capable of capturing complex non-linear patterns and interactions, suitable for high predictive accuracy.
  - **Time Series Forecasting Models (e.g., ARIMA/SARIMA, Prophet, LSTM):** Essential for accurate future price forecasting, as they account for temporal characteristics like trends and seasonality inherent in price data.

5.2. Classification Models (for Price Categorization or Anomaly Detection)

- **Objective:** To classify mp_price into predefined categories (e.g., 'low', 'medium', 'high' price bands) or to identify unusual price fluctuations indicating market disruptions.
- **Applicable Algorithms & Justification:**
    - **For Price Categorization:** Decision Trees or Random Forest Classifiers can effectively assign prices to discrete categories.
    - **For Anomaly Detection:** Isolation Forest or One-Class SVM are unsupervised algorithms ideal for flagging extreme price points that deviate from normal behavior, serving as early warning indicators.

5.3. Clustering Models (for Market/Commodity Segmentation)

- **Objective:** To discover inherent groupings or natural segments within the dataset without predefined labels, such as grouping markets with similar pricing behaviors or commodities with similar price dynamics.
- **Applicable Algorithms & Justification:**
    - **K-Means Clustering:** Useful for partitioning data into a specified number of distinct clusters based on shared characteristics.
    - **DBSCAN:** Capable of finding clusters of varying shapes and identifying noise, which can be valuable for complex market segmentation patterns.

The chosen algorithms will depend on the specific problem definition, desired interpretability, and evaluation performance, leveraging the now-cleaned and analyzed dataset.

6. Conclusion

This report successfully outlines and executes the crucial initial phases of a data analysis project, including data extraction, comprehensive cleaning, and exploratory data analysis (EDA) of the World Food Programme (WFP) Food Prices dataset. We began by meticulously loading and combining raw data, which then underwent rigorous preprocessing to address missing values, correct data types, and engineer relevant features like mp_date. Crucially, outlier treatment was applied to mp_price to enhance data quality and reliability.

The subsequent EDA phase provided invaluable insights into the dataset's characteristics. We gained a clear understanding of price distributions, identified dominant commodities and markets, and observed significant temporal trends in global food prices. The various visualizations generated further illuminated these patterns, making complex relationships easily interpretable.