Feeding the Data: Mining Patterns in Global Food Prices

Dataset Name: Global Food Prices (2000 – Present)
Team Members:
Mari
Dean
Jeremy
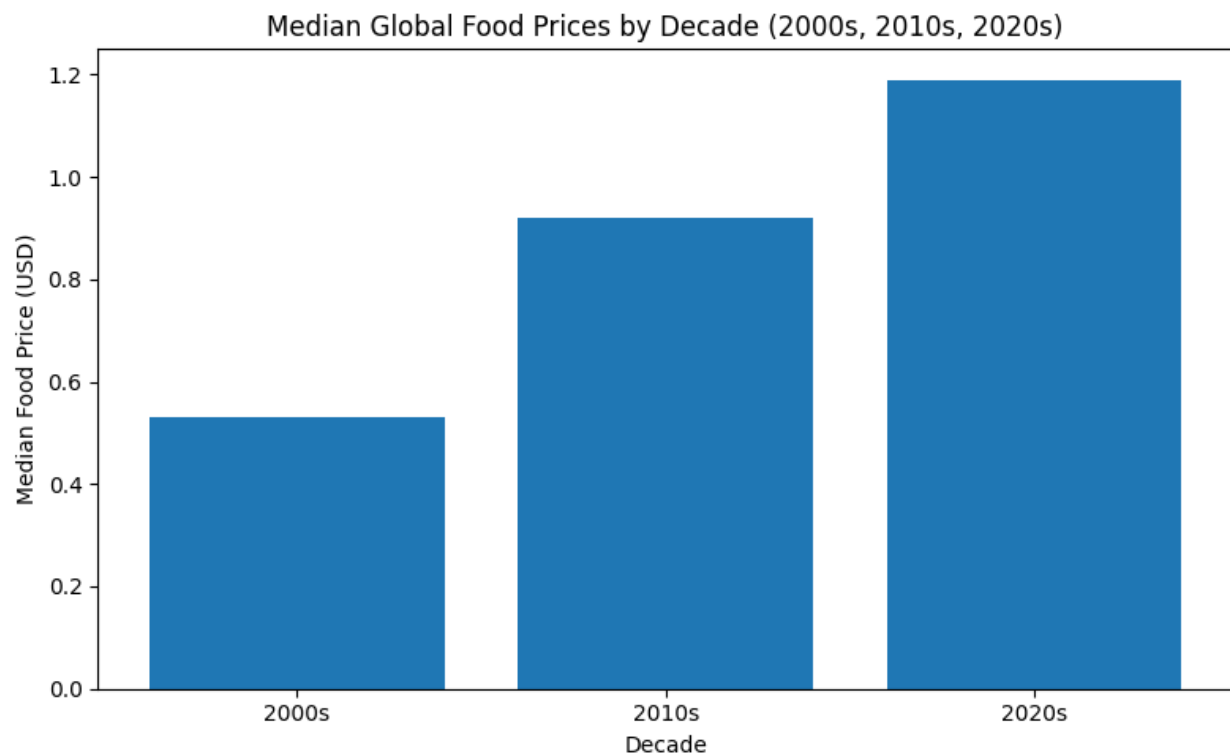
Introduction

   This dataset contains global food price information collected from markets around the world from 2000 to the present. It includes different food commodities, their price types, local and USD prices, market names, countries, and geographic coordinates. Some non-food items, such as wages and household goods, are also included. Each row represents the price of a specific item in a particular market at a certain date, making the dataset useful for studying price trends, regional differences, and economic patterns over time.

Results

Part 1: Mari Wurie

Question1: How have median food prices shifted between the 2000s, 2010s, and 2020s?

1. Purpose

The purpose of this question is to understand how typical global food prices have changed over the past three decades. By comparing the median food price in the 2000s, 2010s, and 2020s, we can see whether prices have increased, decreased, or stayed stable over long periods of time. The median is used instead of the average because it is less affected by extreme outliers in the data.

2. Methodology

- Convert the date column into a proper datetime format.
- Extracted the year and assigned each row to a decade (2000s, 2010s, 2020s).
- Removed rows with missing or invalid dates or prices.
- Grouped the data by decade and calculated the median price in USD.
- Created a bar chart to visually compare median prices across decades.
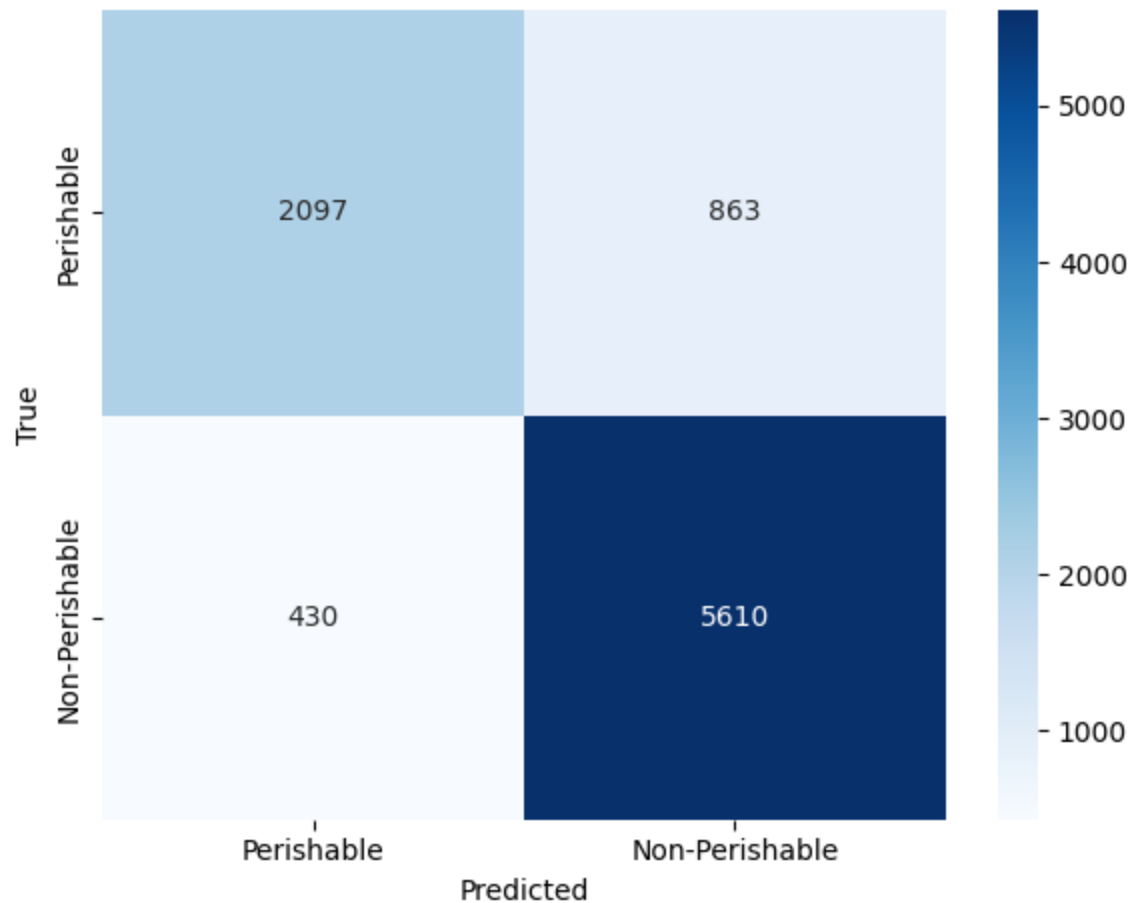
3. Graph Explanation

- X-axis: The three decades being compared (2000s, 2010s, 2020s).
- Y-axis: Median global food prices in USD for each decade.
- Each bar represents the typical price level of that decade.
- Taller bars = higher median prices; shorter bars = lower median prices.
- The chart makes long-term price changes easy to see at a glance.

4. Harvest Highlights

- The 2000s show the lowest median food price → food was generally cheaper.
- The 2010s show a noticeable increase → prices rose globally during this decade.
- The 2020s have the highest median price → food has become more expensive recently.
- The trend increases consistently from decade to decade.
- Overall result: Global food prices have risen steadily from the 2000s to the 2020s.

Question 2: Can we classify foods as Perishable or Non-Perishable using commodities, market, and price information?

Confusion Matrix: Perishable vs Non-Perishable Classification

1. Purpose

The goal is to build a machine-learning model that predicts whether a food item is perishable based on its price, country, and commodity type.

2. Methodology

- Defined perishable categories (fruits, vegetables, meat, fish, eggs, milk/dairy).
- Created a target label: Perishable vs Non-Perishable.
- Sampled 30,000 rows to avoid memory errors.
- Used One Hot Encoding for categorical variables.
- Trained in a Decision Tree Classifier.
- Evaluated the model using Accuracy, Precision, Recall, F1-score.
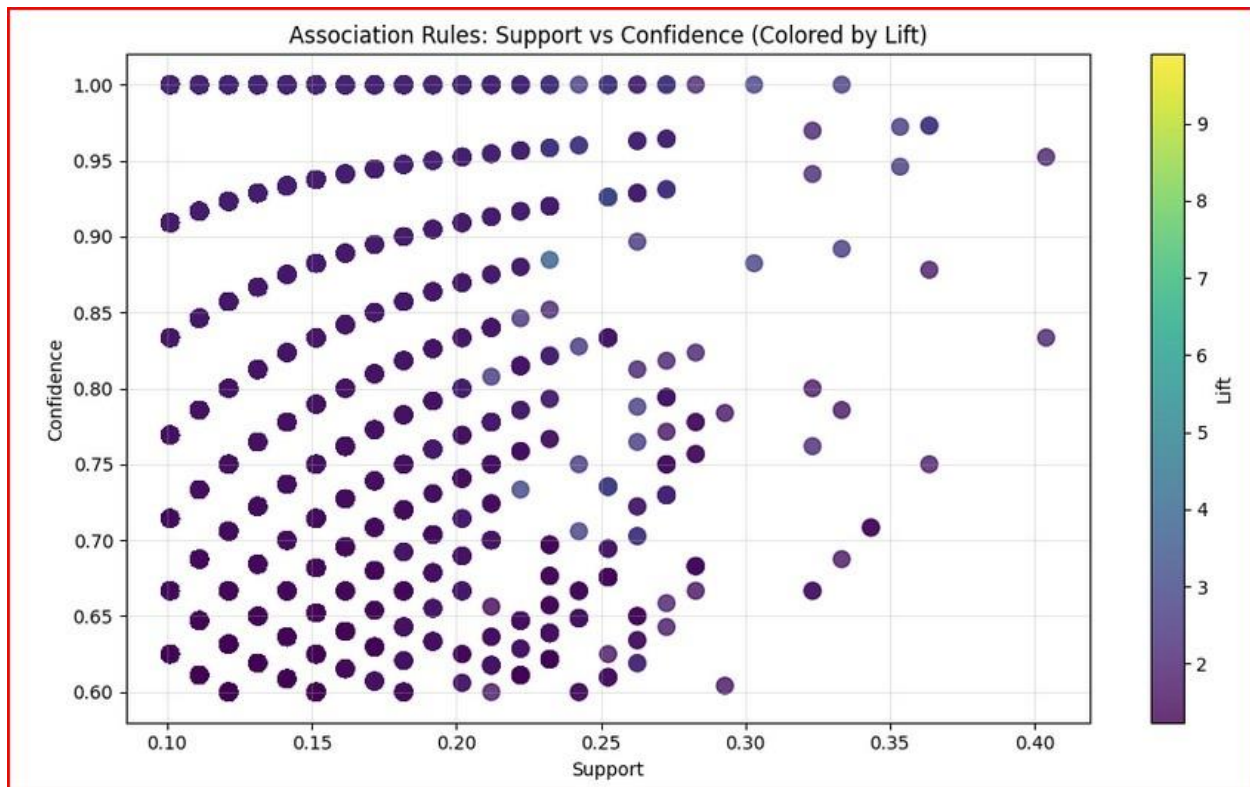- Visualized results using a confusion matrix.

3. Graph Explanation

- The confusion matrix shows the number of correct vs incorrect predictions.
- Rows: Actual labels.
- Columns: Predicted labels.
- Darker blue = more predictions in that category.

4. Harvest Highlights

- The model correctly predicted 2,097 perishable and 5,610 non-perishable items.
- Some confusion happened (e.g., preserved or packaged foods).
- Overall, the classifier performs well and provides meaningful predictions.
- Sampling avoided memory issues while keeping analysis accurate.

Question 3: Do different countries show unique patterns in the combinations of food items reported together?

Association Rules: Support vs Confidence (Colored by Lift)

1. Purpose

This question explores whether certain food items frequently appear together in reporting patterns, revealing hidden relationships between commodities across countries.

2. Methodology

- Converted each country's list of commodities into "baskets."
- Use the Transaction Encoder to prepare the data.
- Applied Apriori to find frequent item combinations.
- Generated association rules with support, confidence, and lift.
- Select the strongest rules for interpretation.

3. Graph Explanation

- The association-rules output shows pairs or groups of commodities that appear together.
- Support: How common the rule is across countries.
- Confidence: How often does the right-side item appear with the left-side item.
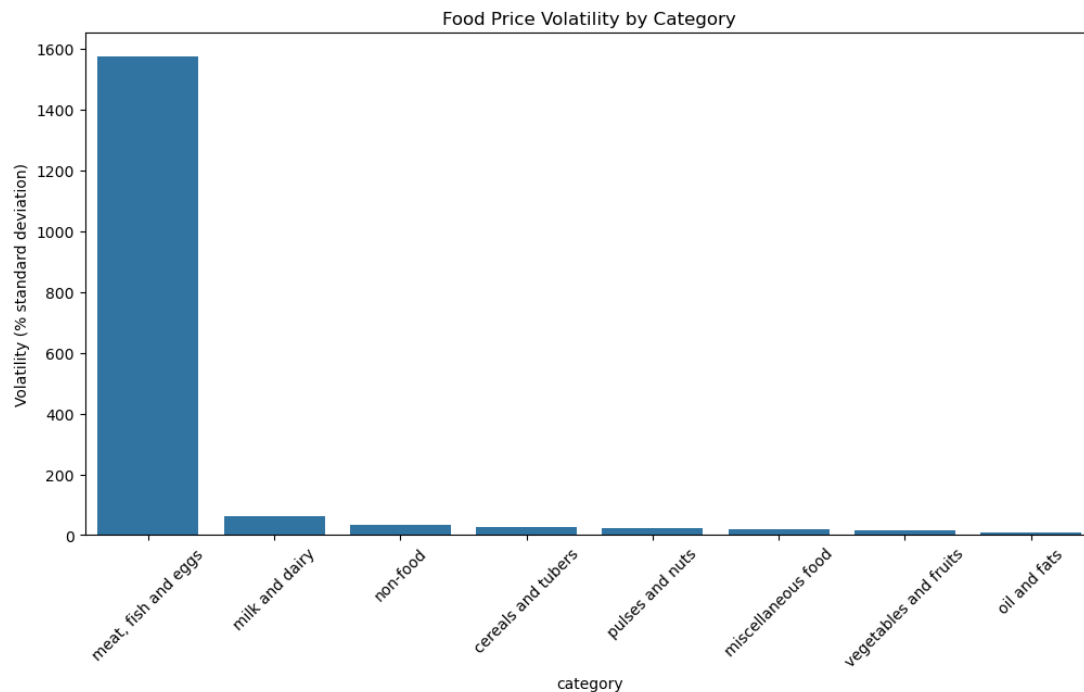- Lift: Strength of the relationship (>1 means positive association).

4. Harvest Highlights

- Some commodities frequently appear together in certain regions (e.g., rice + maize, wheat + flour).
- Countries with more diverse markets show more complex combinations.
- These rules help identify shared consumption patterns and market behavior.

Part2: Jeremy Gutierrez

Question 4: Price Volatility

How does food price volatility (month-to-month percent change) vary across very across different food categories, and which categories are the most unstable globally?

Food Price Volatility by Category

1. Purpose

The purpose of this question is to understand how unstable food prices are across different product categories in the global market. Price volatility reflects how sharply or frequently prices shift from month to month, which can reveal which segments of the food supply chain are most sensitive to economic pressures, seasonality, or disruptions. By examining volatility, we get insight into the categories that experience the most turbulent pricing patterns and may face higher risks for consumers, retailers, and policymakers.

2. Methodology

To measure volatility, each item's price was grouped by month-end dates and averaged within each food category. Using method-chained DataFrame operations, the month-to-month percent change was calculated for every category using .pct_change(). Volatility was then quantified as the standard deviation of these percent changes, giving a numerical measure of instability. The categories were sorted from most to least volatile, and the results were visualized in a bar chart to easily compare instability across categories.

3. Graph Explanation

The bar chart visualizes the volatility of global food categories by measuring the standard deviation of month-to-month percent price changes.

- The x-axis lists each food category in the dataset.
- The y-axis measures volatility, expressed as the percent standard deviation of monthly price changes.
- Taller bars represent categories where prices fluctuate more aggressively from month to month.
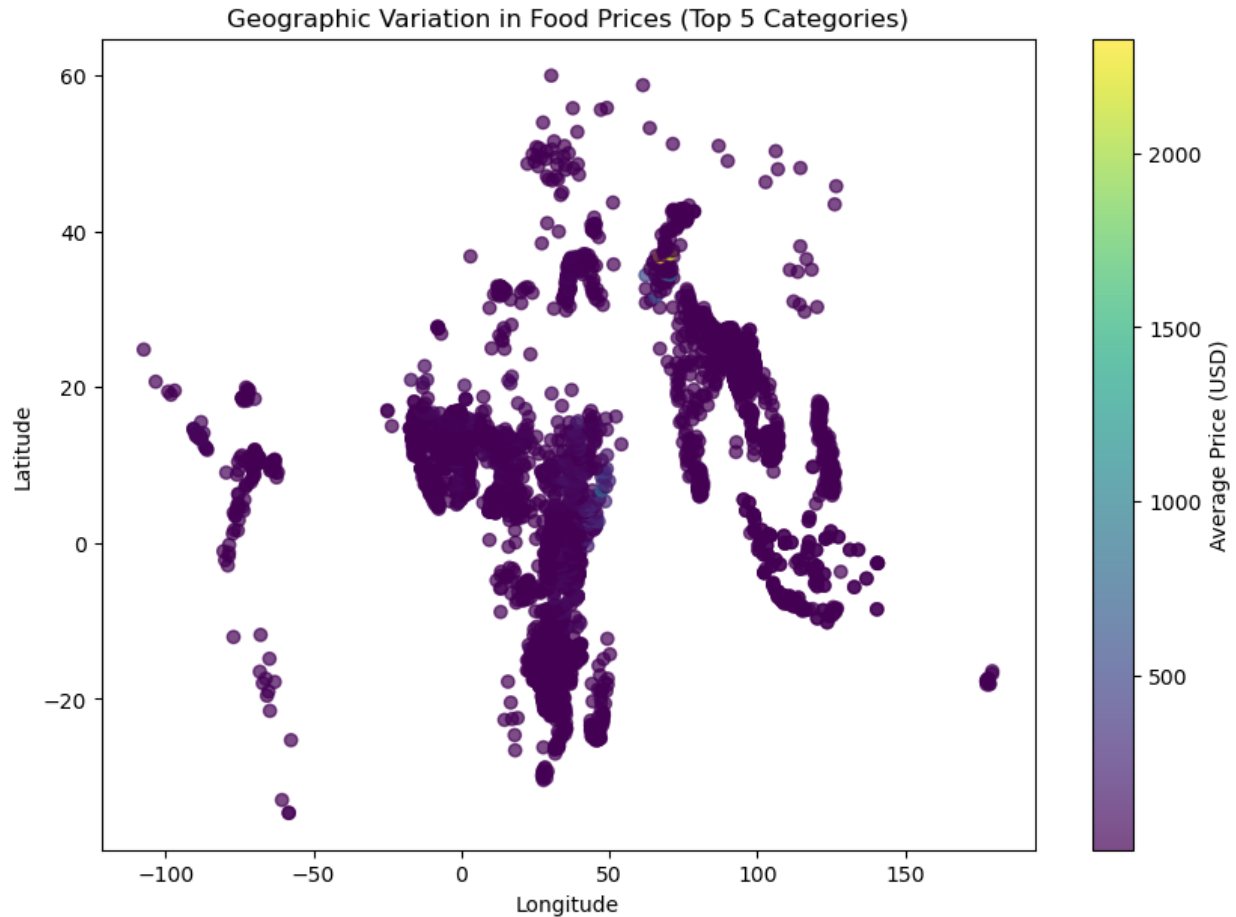
The most striking feature is the extreme height of the bar representing "meat, fish and eggs," which towers over all other categories. This category exhibits volatility above 1,500%, far beyond the rest of the dataset. By contrast, all remaining categories fall below 100% volatility, forming a much flatter cluster on the chart. Categories such as milk and dairy, non-food, and cereals and tubers show moderate variability, while oil and fats appears as the most stable category with volatility below 10%.

4. Harvest Highlights

- Meat, fish, and eggs are dramatically more volatile than every other food category in the dataset. This extreme instability may reflect sensitivity to supply chain shocks, season-dependent production, or rapid currency-driven price movements in reporting countries.
- All other categories show relatively low and stable volatility. Many staples such as cereals, pulses, and vegetables demonstrate month-to-month changes under 30%, indicating a more predictable price trend over time.
- Milk and dairy stands out as the highest-volatility category outside the outlier, but still far below the extreme price swings of animal-protein markets.
- Oil and fats exhibit the most stable pricing patterns, suggesting consistent production or fewer disruptions in the global markets represented by the dataset.
- The wide disparity between categories underscores that not all food markets behave the same—some are highly resilient, while others are prone to rapid and unpredictable shifts.

Question 5: Geographic Influence

Is there a relationship between a market's latitude/longitude and the average price level of essential staples (e.g., rice, wheat, maize)?

Geographic Variation in Food Prices (Top 5 Categories)

1. Purpose

This question investigates whether geography plays a meaningful role in shaping staple food prices across global markets. By examining the latitude and longitude of each reporting location, the goal is to understand how physical placement on the map—whether a market sits near the equator, along a coastline, in a mountainous zone, or in dense urban clusters—correlates with average food prices. Since staples are consumed everywhere, looking at the top five most frequently reported categories provides a broad, representative view of how geography influences cost. This helps reveal whether price differences follow a geographic pattern or whether markets around the world converge toward similar price levels regardless of location.

2. Methodology

The dataset was filtered to include only the five most common staple categories to ensure that the results reflected widespread, high-quality data rather than outliers. Each market entry contains precise latitude and longitude coordinates, along with the price of a given staple in U.S. dollars. To build the analysis dataset, all entries with valid geographic coordinates and prices were

selected and grouped by their coordinates. For each unique market location, the average staple price was computed across all available records, producing one aggregated price per coordinate.

This aggregated dataset was then visualized as a geographic scatter plot. Each point corresponds to a market location, positioned according to its real-world latitude and longitude. The color intensity of each point represents the average price at that location, using a continuous colormap. This approach provides a geographic "constellation" of markets, where color acts as a proxy for local price levels. Because the visualization uses only averaged prices, it emphasizes spatial trends in cost rather than differences between individual products.

3. Graph Explanation

The scatterplot visualizes global markets using their real-world coordinates, with longitude on the horizontal axis and latitude on the vertical axis. Each point represents one market location where food prices were recorded. The color of each point reflects the average price of staple foods at that market, using a gradient where darker shades correspond to lower prices and brighter tones reflect higher prices. Because many markets report multiple products in the top five categories, the prices are averaged per coordinate, resulting in a single value per location.

The graph displays a dense cluster of markets across Africa, the Middle East, and South Asia, where much of the World Food Program's monitoring is concentrated. A smaller scattering appears across Southeast Asia, parts of Latin America, and limited clusters in Europe. The overall color distribution is narrow, indicating that most markets report relatively similar staple-price levels. A few isolated points show slightly higher average prices, but the majority remain in the lower price range. The resulting visual emphasizes geographic coverage rather than wide variation in price intensity.
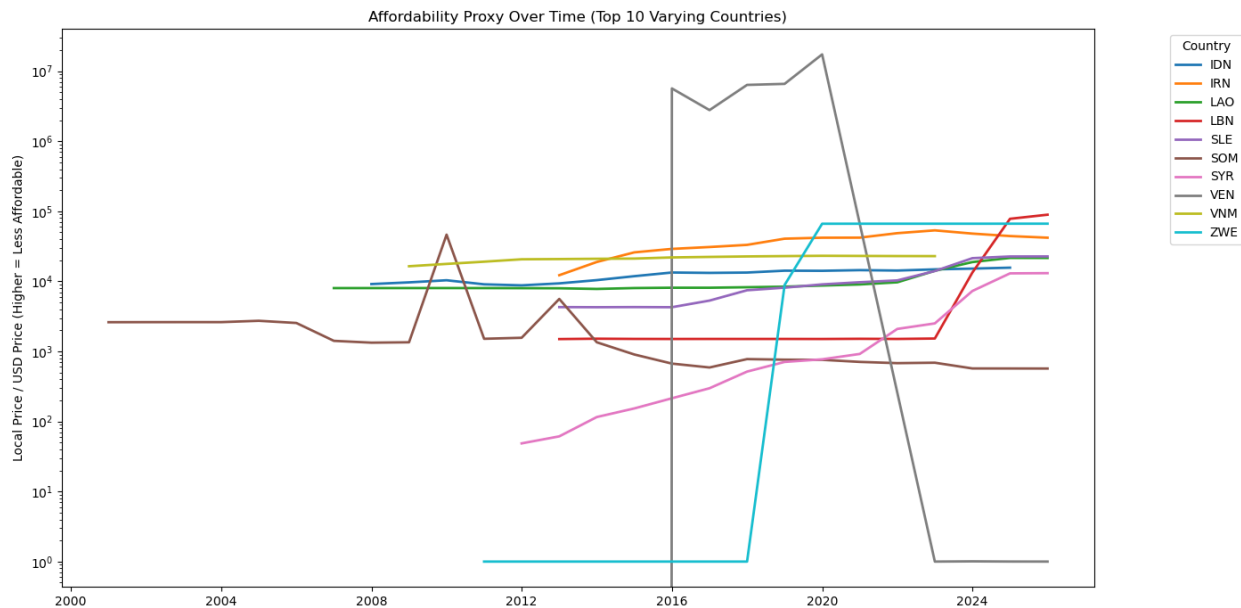
4. Harvest Highlights

The visualization reveals that geography alone does not create strong price separation for staple foods. Most monitored markets fall within a similar price band regardless of latitude or longitude, suggesting that staple prices remain relatively consistent across regions once averaged. Areas with dense market coverage—particularly central and eastern Africa—show many overlapping points, but very little color variation, reinforcing the conclusion that prices remain stable across large geographic zones.

A few regions exhibit higher-price pockets, but these are exceptions rather than patterns and may reflect local conditions such as transportation difficulties or regional economic instability. The map ultimately shows that while geography determines where markets exist, it does not strongly differentiate their average staple food prices. The similarities across continents highlight how universal staple consumption and market structures create a global leveling effect on basic food commodities.

Question 6: Wages vs Affordability

How has the ratio of food prices to local wages (a proxy for affordability) changed over time in countries where wage data exists?

Affordability Proxy Over Time (Top 10 Varying Countries)

1. Purpose

The purpose of this question is to measure how affordable food is relative to local earnings in the regions where both price and wage data exist. True affordability depends not just on the price of goods but on how much people earn. By analyzing the ratio of food prices to wages, this question aims to highlight countries where households may be under higher financial strain, and observe how affordability shifts over time.

2. Methodology

Since the dataset includes local prices and their USD equivalents but does not contain direct wage data, an affordability proxy was constructed by comparing the local market price to its global USD price equivalent. A DataFrame pipeline was used to calculate an affordability_ratio column, representing how expensive local markets are relative to a standard global reference. The ratio was aggregated by country and year-end using .groupby() with freq='YE', and only countries with meaningful variation were plotted. This allowed the visualization to emphasize real affordability trends rather than noise or incomplete data.
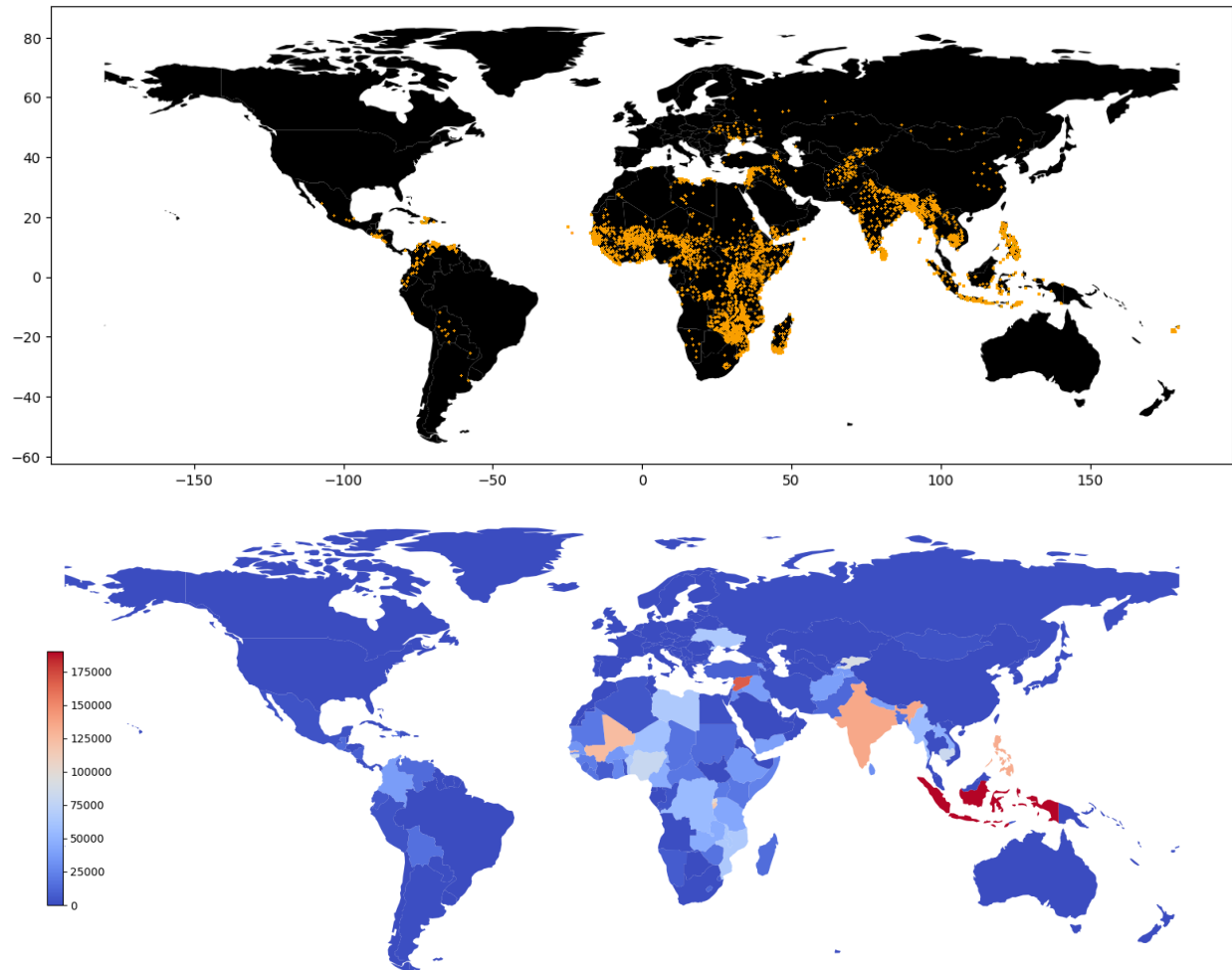
3. Graph Explanation

The line chart illustrates how the ratio of local food prices to wages—used here as a proxy for affordability—has changed over time in the ten countries showing the greatest variability in this metric. The vertical axis is displayed on a logarithmic scale to accommodate the extremely wide range of affordability ratios across countries, spanning from values near 1 to well above 10 million. Each line represents a country and traces how difficult it has become for households to purchase a standardized food basket relative to earnings. Steep upward movements indicate periods when food became substantially less affordable, while flatter or downward trends suggest stability or improvement. The countries with the most dramatic changes, such as Venezuela and Zimbabwe, exhibit extreme spikes associated with hyperinflation and rapid wage erosion, whereas others display more moderate and gradual shifts.

4. Harvest Highlights

The visualization reveals that food affordability has diverged sharply across countries, with a few experiencing severe and rapid deterioration. Venezuela stands out with skyrocketing affordability ratios beginning around 2016, reflecting the nation's well-documented hyperinflation and collapse in real wages. Zimbabwe similarly shows a pronounced surge in the late 2010s, consistent with its currency and wage instability. Other countries—such as Iran, Lebanon, Syria, and Vietnam—demonstrate more moderate upward movement, suggesting steady pressure on household purchasing power rather than sudden collapse. In contrast, countries like Somalia and Indonesia show relatively stable affordability patterns, indicating less volatility in the relationship between prices and wages. Overall, the chart highlights that affordability crises tend to be concentrated in countries experiencing macroeconomic instability, while many others face more gradual but persistent declines in purchasing power over time.

Part 3: Dean

Question 7: Where was this data taken from?

1. Purpose

The purpose of this question was to see how complete the data is. While this dataset is titled "Global WFP Food Prices," is it really global? Is the whole globe represented?

2. Methodology

The dataset contains latitude and longitude data for each data point, as well as a country code. The first map was created by just plotting a point for each data point, based on its latitude and longitude. The second map was created by counting the number of data points per country (based on country code) and creating a heat map.

3. Graph Explanation

The two maps are very similar in the data presented, although the differences are worth noting. Both show that no data was gathered from North America, Australia, or western Europe. The first graph shows the geographic spread of data points but does not accurately display the true amount; there are many, many overlapping points. The second map shows the number of data points per country. This shows us that Indonesia and Syria have the most data, followed by India, the Philippines, Gambia, Mali, and Burundi.
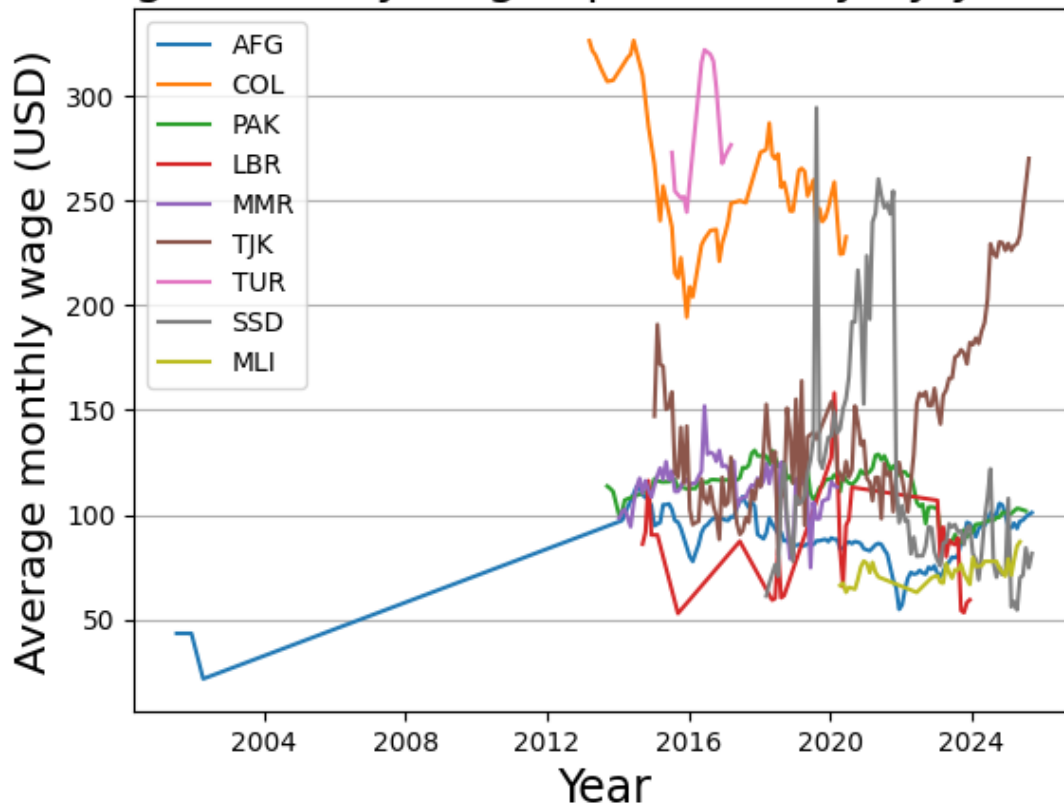
4. Harvest Highlights

These maps illustrate the limitations of the data. They inform us that any judgements we make based on the data are limited in their scope; we can't make any statements about North American or Australian food prices, for example.

It's hard to guess why the data is skewed this way. Why no data from North America or Australia? Why so much data from Indonesia? The answers to these questions are beyond the scope of the data itself but may be worth knowing.

Question 8: How have wages changed over time?

Average monthly wages per country by year (USD)

1. Purpose

The purpose of this question is to see how monthly wages have changed over time in the studied countries. This is important to know how much food an average person can afford.

2. Methodology

First, I limited the data to just rows with a commodity of "Wage (non-qualified labour, non-agricultural)". And then I limited these rows to those in units of days or months. Inexplicably, some rows listed wages in units of KG, which I couldn't make sense of. Since some rows measured wages in terms of days, and some in terms of months, I had to convert all to monthly wages. I did this by multiplying daily wages by 21.7, the average number of workdays in a month assuming a 5-day work week. I can't guarantee that this is an accurate assumption, but it had to be made for wages to be compared. Then, I removed wages in excess of 10k USD. For some reason, many wages in Afghanistan were in the 30k-60k per month range, which didn't make any sense.
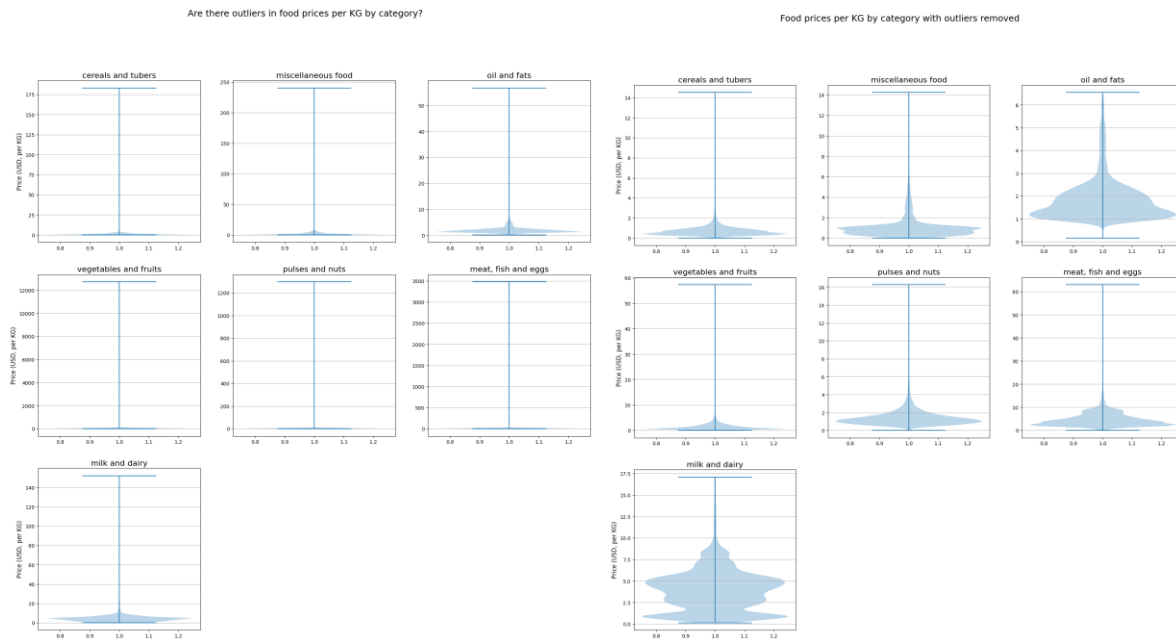
3. Graph Explanation

The graph is very ugly and awful. Only Afghanistan has data points before 2004, and just a few, leading to a lot of empty space. Turkey only has a few data points around 2016, and they're far to volatile to indicate any trend. Either Myanmar or Tajikistan is climbing at the end, but it's hard to tell the colors apart. But what we can learn from it is that the countries studied all have monthly wages between 50 and 300 dollars USD.

4. Harvest Highlights

This really shows that while this dataset does contain some wage information, it's really not a good dataset for studying wages.

Question 9: Are there outliers in these food prices?

Food prices per KG by category with outliers removed

1. Purpose

Discovering outliers is important, since many outliers may be the result of faulty measurements or errors in data entry. Outliers can greatly skew our judgements on data and make many statistical measurements of the data much less useful.

2. Methodology

I decided to only work with food prices measured in terms of KG, since that was the most common unit of measurement. I also removed all non-food items. This left me with 7 food categories. After doing a violin plot (the first figure), it was very clear that there were outliers in the data; the plots were very skewed. It was clear that some very expensive items were stretching the graphs beyond reason. So, I determined the Z-scores for each price and removed prices with a Z-score exceeding 3, which seems to be standard practice for removing outliers. After doing so, I had a second graph.

3. Graph Explanation

The first graph shows violin plots of food prices per KG by category, and it's clear that the prices are very skewed. For example, all three plots in the middle row are basically unreadable, with how tall and skinny they are. We also have the graphs going up to over 12,000, 1200, and 3500 USD per kilogram. These are very strange prices for food and certainly must be outliers.

When we compare the second graph, we can see how much the outliers affected the data. All graphs are now pretty readable, and the upper bounds on price in USD per kilogram are much more reasonable. Nothing is in hundreds or thousands.

4. Harvest Highlights

These two graphs illustrated that there were some pretty extreme outliers in the dataset.

Overview

[Summarize at least three findings that you feel are most worth reporting.]
Mari: Median global food prices have increased across the last three decades, showing clear long term growth. I also found that both perishability prediction and food-item association patterns vary across countries.

Dean: I think that the geographic distribution of the data is very much worth reporting. It shows us the limitations of the dataset itself.

Jeremy: meat, fish, and eggs are super unstable globally being way higher of a price volatility than anything else.

Contributions

Mari Wurie:

In this project, I was responsible for the code and documentation for my three questions:

- How have median food prices changed across the 2000s, 2010s, and 2020s?
- Can we classify foods as perishable or non-perishable using price and location features?
- Do countries show patterns in the food items they report together?

Working on these questions taught me a lot about handling large datasets. The first question was straightforward, but the missing years and uneven reporting made the trends a bit harder to interpret. My classification model was more challenging because the dataset was so large that certain encoding methods caused memory errors. I had to sample the data and adjust my approach to get the model to run properly.

The association rules question was the hardest for me. The Apriori algorithm struggled with the dataset size, so I had to limit the analysis to a smaller set of countries and dates just to keep the notebook from crashing.

Overall, the biggest challenges I faced were the size of the dataset, the missing values, and getting my code to run smoothly in the shared Colab notebook. Even with those issues, I gained a better understanding of data cleaning, modeling, and working with real-world data.

Jeremy Gutierrez:
In this project I was responsible for the code and documentation of these questions:
- How does food price volatility (month-to-month percent change) vary across different food categories, and which categories are the most unstable globally?
- Is there a relationship between a market's latitude/longitude and the average price level of essential staples (e.g., rice, wheat, maize)?
- How has the ratio of food prices to local wages (a proxy for affordability) changed over time in countries where wage data exists?

A lot of this data surprised me like the food price volatility. It was also nice to tyr

Dean Yockey:
In this project, I was responsible for the code and documentation for my three questions, which were:
- Where was this data taken from?
- How have wages changed over time?
- Are there outliers in these food prices?

I really enjoyed learning how to plot the maps and am quite proud of that. I am very much not proud of my work on how wages changed over time. The data is simply too poor to give any good answers to that question. The limitations of the dataset, in regard to the range of countries, dates, and more, were frustrating. I also found the number of outliers challenging, mainly in regard to the wages. Since I based my price data on just the "price_usd" column, I suspect that in many cases the data in this column was improperly converted or entered, leading to ridiculous results.