

# Data Collection Plan & Raw Data Sources Identification Report

Project Title: Global Food Production Trends and Analysis (1961–2023)

## 1. Purpose of Data Collection

The purpose of this plan is to identify, acquire, and prepare reliable data sources to analyze global food production patterns from 1961 to 2023. The collected data will be used to develop **Power BI dashboards** for trend analysis, country-level comparisons, and commodity-wise insights.

## 2. Data Collection Plan

### 2.1 Data Requirements

To achieve the project objectives, the following types of data are required:

- **Country-Level Information** – Entities, regions, and income group classifications
- **Time Series Data** – Annual production from **1961 to 2023**
- **Commodities** – Food items (100+) including *Cereals, Fruits, Oilseeds, Root Crops, Cash Crops, Beverages*
- **Production Volumes** – Reported in tonnes, convertible to billion tonnes for readability

### 2.2 Data Source Selection Criteria

- **Authenticity** – Data must come from trusted global organizations or curated platforms
- **Coverage** – Must span multiple decades (1961–2023) and multiple commodities
- **Granularity** – Country-wise and item-wise detail
- **Accessibility** – Open and publicly available for analysis

## 3. Raw Data Sources

- **Source:** [Kaggle – World Food Production](#)
- **Timeframe:** 1961–2023
- **Records:** Country-wise annual production volumes for 100+ commodities
- **Key Attributes:**
  - Country/Entity
  - Year
  - Item (Commodity)
  - Production (in tonnes)

## 4. Data Acquisition & Storage Plan

- **Acquisition Method:** Direct download from Kaggle (CSV format)
- **Storage:**
  - Initial exploration in Microsoft Excel
  - Integration into Power BI for cleaning, transformation, and visualization
- **Version Control:** Maintain raw and cleaned datasets separately to preserve original integrity

## 5. Data Preprocessing Strategy

- **Data Cleaning:** Remove null values, handle duplicates, correct inconsistent entity names
- **Filtering:** Exclude non-food or irrelevant items
- **Categorization:** Create grouped categories (*Cereals, Fruits, Oilseeds, Root Crops, Cash Crops*)
- **Unit Standardization:** Convert production values into **billions of tonnes** for uniformity
- **Validation:** Cross-check sample data points against FAOSTAT for accuracy

## 6. Deliverables

- **Interactive Power BI Dashboards** (2 views)
- **Project Documentation** (Problem Statement, Proposal, Planning Report, Insights)
- **Insights Report** summarizing key findings

## 7. Expected Outcomes

- A **clean, reliable, and structured dataset** ready for Power BI analysis.
- Coverage of **global food production trends** across 60+ years and 100+ commodities.
- Robust foundation for building **interactive dashboards** and deriving **meaningful insight**.