Data Preprocessing Report

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

1. Objective of Preprocessing

The raw dataset obtained from Kaggle/FAOSTAT required preprocessing to ensure consistency, reliability, and usability in Power BI. The goal was to transform the raw agricultural production data into a structured, analysis-ready format that supports interactive dashboards and accurate KPIs.

2. Preprocessing Steps

Step 1: Data Import & Inspection

- Imported dataset (CSV) into **Power Query Editor** in Power BI.
- Inspected schema: columns included Entity, Year, Item, and Production (tonnes).
- Verified data types (e.g., Year as integer, Production as decimal).

Step 2: Data Cleaning

- **Null Removal:** Eliminated rows with missing or null Entity or Item.
- **Irrelevant Commodities:** Filtered out non-food items and duplicate commodity codes.
- **Entity Name Standardization:** Harmonized country/entity names for consistency (e.g., “USA” → “United States”).
- **Date Formatting:** Ensured Year column values strictly ranged between **1961–2023**.

Step 3: Feature Engineering

- **Commodity Categorization:** Created a new column Commodity Category grouping items into:
    - *Cereals, Fruits, Root Crops, Oilseeds, Cash Crops, Beverages, Others*.
- **Unit Conversion:** Converted production values from tonnes → **billion tonnes** for improved readability.
- **Derived Columns:** Added calculated columns for year-on-year growth and percentage contributions.
- **DAX Measures:** Defined custom measures to support KPIs and dynamic visuals:
    - *Total Production*
    - *Commodity Share %*
    - *Top Producer by Year*
    - *Growth Rate*

Step 4: Data Transformation for Visualization

- Reshaped tables to support comparative visuals (e.g., pivoted Item for bar charts).
- Created **hierarchical relationships** (Entity → Commodity → Year).
- Ensured data was optimized for slicers and filters (Year, Commodity, Country).

Step 5: Data Validation

- Cross-checked aggregate totals against FAOSTAT reference values.
- Validated that derived KPIs matched expected trends (e.g., Sugarcane as the highest-producing commodity globally).
- Ensured no duplicates or negative values remained in the dataset.

Overall Quality: High – Dataset is suitable for advanced analysis and visualization after minor preprocessing.

3. Outcome of Preprocessing

After preprocessing, the dataset was transformed into a **clean, structured, and analysis-ready format**, enabling:

- Reliable time-series analysis (1961–2023).
- Commodity- and country-level comparisons.
- Accurate KPI tracking and dynamic Power BI dashboards.

Final Status: Data preprocessing successfully completed; dataset is ready for advanced visualization and trend analysis.