

Data Quality Report

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

1. Dataset Overview

- Total Records: ~60,000+ rows (covering 1961–2023)
- Attributes:
 - Entity (Country/Region)
 - Year
 - Item (Commodity)
 - Production (tonnes)

The dataset captures annual production volumes of over 100 food commodities across multiple countries and regions over six decades.

2. Data Quality Dimensions

2.1 Accuracy

- The dataset originates from **FAOSTAT**, ensuring high reliability and accuracy of reported figures.
- Spot checks with official FAO records confirmed consistency in production values for major commodities like **Sugarcane, Wheat, and Rice**.

Status: **High Accuracy**

2.2 Completeness

- All years from 1961 to 2023 are covered.
- Some missing entries for certain countries/commodities (expected due to lack of reporting in earlier years).
- Null values primarily appeared in production quantities for lesser-reported crops.

Status: **Mostly Complete** (with minimal gaps)

2.3 Consistency

- Duplicate entries were not observed in the raw Kaggle dataset.
- Country/entity names had variations (e.g., “United States” vs. “USA” in FAO sources). Standardization was required.
- Units of production were consistent (all values reported in **tonnes**).

Status: **Consistent with minor name variations** (resolved during cleaning)

2.4 Timeliness

- Dataset includes data up to **2023**, making it current and relevant.
- Annual granularity supports long-term trend analysis.

Status: **Up-to-date**

2.5 Validity

- All Year values fall within the valid range (1961–2023).
- Production values were numeric and logically aligned (no negative or unrealistic numbers).
- Outliers (e.g., extreme spikes in production) reflect genuine agricultural trends (e.g., sharp increases in maize post-1980s).

Status: **Valid**

2.6 Interpretability

- Attributes are clearly defined (Entity, Year, Item, Production).
- Commodity names are human-readable and mapped to FAO standard categories.
- Further interpretability improved by creating Commodity Categories (*Cereals, Fruits, Root Crops, Oilseeds, Cash Crops*).

Status: **Highly Interpretable**

3. Data Quality Issues Identified & Resolution

Issue	Description	Resolution
Missing Values	Some countries lacked production data for certain crops/years	Removed nulls or replaced with 0 where logically valid
Inconsistent Country Names	Variations in country/entity labels	Standardized using Power Query transformations
Large Numbers Difficult to Read	Production reported in tonnes (e.g., billions of tonnes difficult to interpret)	Converted to billions of tonnes for readability
Irrelevant Items	Non-food or duplicate entries included	Filtered out during preprocessing

4. Overall Data Quality Assessment

Dimension	Rating	Notes
Accuracy	★★★★☆	Verified against FAO records
Completeness	★★★★☆	Minor gaps in some country/commodity data
Consistency	★★★★☆	Standardization required for entity names
Timeliness	★★★★★	Includes data till 2023
Validity	★★★★★	No invalid/negative values
Interpretability	★★★★★	Clear attributes, categorized for analysis

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

5. Conclusion

The **World Food Production dataset (1961–2023)** demonstrates strong reliability, completeness, and consistency. After resolving minor issues through cleaning (null handling, entity standardization, unit conversion), the dataset is of **sufficient quality for building interactive Power BI dashboards** that provide meaningful insights into global food production trends.