**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 24 June 2025 |
| Team ID | xxxxxx |
| Project Title | Global Food Production Trends and Analysis (1961–2023) using power bi |
| Maximum Marks | 3 Marks |

**Data Quality Report Template**

The Data Quality Report will summarize potential issues within the food production datasets. Key dimensions include completeness, consistency, accuracy, timeliness, and validity. Completeness issues may arise due to missing data for certain countries or years, particularly in conflict-affected or small nations. Consistency challenges occur when country names, commodity codes, or measurement units differ across sources. Accuracy must be validated by cross-checking with secondary sources to avoid reporting errors. Timeliness concerns are limited to the latest updates, as official datasets may lag behind current years. Each issue will be assigned a severity level, and resolution plans will involve standardization, cleaning, or transparent documentation to ensure integrity for analysis.

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Source** | **Data Quality Issue** | **Severity** | **Resolution Plan** |
| World Food Production (Kaggle) | Inconsistent country names and codes across years (e.g., “USA”, “United States”, “U.S.”). | High | |  | | --- | | Standardize using ISO3 country codes and apply uniform naming conventions with a mapping dictionary. |  |  | | --- | |  | |
| |  | | --- | |  |  |  | | --- | | World Food  Production (Kaggle) | | |  | | --- | | Missing values for certain countries/commodities in  some years (1961–2023). | | Moderate | |  | | --- | | Use interpolation for time-series continuity  or leave nulls documented transparently  to avoid bias. |  |  | | --- | |  | |
| |  | | --- | |  |  |  | | --- | | World Food  Production (Kaggle) | | |  | | --- | | Commodity naming  inconsistencies  (e.g., “Corn” vs. “Maize”). |  |  | | --- | |  | | Moderate | |  | | --- | | Map all commodities to FAO/standardized  codes before merging and analysis. |  |  | | --- | |  | |
| |  | | --- | |  |  |  | | --- | | World Food  Production (Kaggle) | | |  | | --- | | Derived yield values may  have rounding/precision  discrepancies. |  |  | | --- | |  | | Low | |  | | --- | | Recalculate yield consistently as  Production ÷ Area to maintain accuracy |  |  | | --- | |  | |
| |  | | --- | |  |  |  | | --- | | World Food  Production (Kaggle) | | |  | | --- | | Large dataset size may  slow visualization tools  like Power BI. | | Low | Optimize by aggregating at regional/decade level, using query folding, or preprocessed CSV extracts. |