**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 18 August 2025 |
| Name | Mohammed Farhan |
| Project Title | India Agriculture Crop Production Analysis (1997–2021) Using Tableau |
| Maximum Marks | 3 Marks |

**Data Quality Report Template**

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Source** | **Data Quality Issue** | **Severity** | **Resolution Plan** |
| Main Production Dataset (1997–2021) | Missing values in Area/Production/Yield across years and states | High | Impute numerics by (State, Crop) median; recompute Yield = Production ÷ Area where possible; flag unresolved records for review. |
| Main Production Dataset (1997–2021) | Inconsistent crop names and aliases (e.g., “Ground nut”, “Arhar/Tur”, “Bajra (Pearl millet)”) | High | Apply a canonical crop dictionary; trim/normalize case; maintain a mapping table (raw→canonical); enforce uniqueness constraints in ETL. |
| Main Production Dataset (1997–2021) | Inconsistent state names and administrative changes (e.g., Orissa→Odisha; Telangana formation in 2014) | High | Standardize names to current official list; add State\_Code; annotate pre/post-bifurcation periods; avoid retroactive splitting of pre-2014 Andhra Pradesh. |
| Main Production Dataset (1997–2021) | Duplicate rows on (Year, State, Crop) composite key | Moderate | Create composite primary key; drop exact duplicates; for near-duplicates, retain the record with complete fields; log and audit drops. |
|  |  |  |  |
| Main Production Dataset (1997–2021) | Numbers stored as text with commas/whitespace | Moderate | Strip non-numeric characters; convert to numeric types; assert no conversion failures; unit tests in preprocessing pipeline. |
| Main Production Dataset (1997–2021) | Unit inconsistencies (tonnes vs thousand tonnes) | High | Normalize all quantities to metric tonnes; apply scale factors based on metadata; add a Unit column and verify aggregates against reference totals. |
| Main Production Dataset (1997–2021) | Zero/negative Area with positive Production causing invalid Yield | High | Set invalid Area to null; recompute Yield; if unresolved, mark Yield null and exclude from yield analytics; add rule-based QC checks. |
| Main Production Dataset (1997–2021) | Outliers/spikes (implausible YoY changes) within State–Crop groups | Moderate | Detect via IQR/z-score over rolling window; winsorize or flag for manual verification; do not auto-correct unless confirmed by source. |
| Main Production Dataset (1997–2021) | Year outside target range or stored as string | Low | Filter to 1997–2021; cast Year to integer; enforce range assertion in ETL. |
| Main Production Dataset (1997–2021) | Encoding and punctuation issues (smart quotes, en-dashes, hidden control characters) | Low | Normalize to UTF-8; replace smart quotes/hyphens; strip control characters and trailing spaces. |
| Main Production Dataset (1997–2021) | Null markers like NA/—/-/Not Available treated as strings | Moderate | Standardize all such markers to actual nulls before type conversion; maintain a list of recognized null tokens. |
| Main Production Dataset (1997– | Column header/schema inconsistencies across multiple CSV files | Moderate | Adopt a canonical schema with exact column names; apply a renaming map during ingestion; validate schema with automated |
| 2021) |  |  | checks. |
| Main Production Dataset (1997–2021) | Category hierarchy double-counting (e.g., Pulses total vs individual pulse crops) | Moderate | Introduce crop hierarchy (Category → Crop); avoid mixing totals with components in the same aggregation; document aggregation rules in report. |
| Main Production Dataset (1997–2021) | Mixed seasonal/annual records (Kharif/Rabi vs annual) | Low | Confirm frequency in metadata; if seasonal present, separate into distinct views or aggregate to annual consistently. |
|  |  |  |  |
| Processed\_Agri\_Data.csv | Post-cleaning verification: mismatched derived Yield after transformations | Low | Recompute Yield deterministically; cross-check with original fields; add validation rule: |Yield - Production/Area| < ε. |
|  |  |  |  |
| Processed\_Agri\_Data.csv | Residual duplicates after joins or reshaping | Low | Re-assert uniqueness on (Year, State\_Code, Crop\_ID); run duplicate detection at the end of pipeline; export exception list. |
|  |  |  |  |
| Lookup – State Mapping | Missing or ambiguous mappings for legacy state names | Moderate | Complete mapping table; add valid-from/valid-to dates; default to official MoHUA/GoI names; version the lookup file. |
|  |  |  |  |
| Lookup – Crop Mapping | Synonyms and misspellings not covered initially | Low | Iteratively expand mapping dictionary from encountered variants; add tests to prevent regressions. |
|  |  |  |  |
| Validation – Range/Business Rules | Area or Production values exceeding plausible maxima for a state | High | Add upper-bound checks using historical maxima per State–Crop; flag breaches for manual review before publishing dashboards. |