Data Hygiene Checklist

- **Source Identification**: Are all data sources cataloged?
- **Ownership**: Is each dataset assigned to a responsible owner?

- **Completeness**: Are all required fields populated?
- **Consistency**: Are formats standardized (dates, currencies, IDs)?
- **Accuracy**: Are values correct and validated?
- **Duplicates**: Are duplicate records identified and removed?
- **Timeliness**: Is data updated regularly?
- **Compliance**: Is sensitive data flagged and protected?
- **Bias Review**: Are demographic groups fairly represented?
- **Documentation**: Is there a data dictionary and schema validation?

Helpful Metrics

- **Data Completeness (DC)** - *Measures how complete datasets are.*
  DC = (Complete Records ÷ Total Records) × 100

- **Data Accuracy (DA)** - *Measures correctness.*
  DA = (Correct Values ÷ Total Values) × 100

- **Duplicate Rate (DR)** - *Measures redundancy.*
  DR = (Duplicate Records ÷ Total Records) × 100

- **Timeliness (T)** - *Measures freshness of data.*
  T = (Records Updated ÷ Total Records) × 100

- **Bias Index (BI)** - *Measures fairness across variables.*
  BI = (Sample of Group A – Sample of Group B) ÷ Total Population

- **Data Usability Score (DUS)** - *Measures overall readiness for AI use..*

  DUS = (Completeness + Accuracy + Timeliness – Duplicate Rate – Bias Penalty) ÷ 5

Path to Use in Organizations

1. Run the checklist across all critical data sources.
2. Calculate metrics to identify weak points.
3. Prioritize remediation (e.g., deduplication, standardization, compliance checks).
4. Partner with your AI Agency to build automated data pipelines with hygiene checks baked in.