Task D — Implement Data Cleaning and Standardization Modules

Project: Real Estate Sentiment & Market Insight

1. Objective

This task defines data cleaning and standardization modules that apply normalization logic to raw ingested data. The process ensures all data is consistent, reliable, and ready for valuation, analytics, and scoring.

2. Input Data

The cleaning and standardization modules process raw datasets from the ingestion layer, including:

- Property data
- Market data
- Foreclosure records
- Auction data
- Comparable sales data

These datasets may contain missing values, inconsistent formats, duplicate records, or invalid entries.

3. Data Parsing and Type Standardization

The first step parses raw fields into consistent data types:

- Convert numeric fields (prices, square footage, bedrooms) to numeric formats
- Convert date fields to standard ISO format (YYYY-MM-DD)
- Convert boolean or categorical fields to standardized values

Records that fail type conversion are flagged for review.

4. Data Cleaning Rules

4.1 Missing Values

- Required identifiers (property_id, market_id) must be present
- Records missing required identifiers are quarantined
- Non-critical missing fields are allowed but flagged

4.2 Duplicate Records

- Duplicates are detected using stable keys (e.g., property_id + event_date + event_type)
- The most complete or latest record is retained

4.3 Outlier Detection

- Values significantly outside market norms (prices, square footage) are flagged
- Outliers are retained but marked to avoid data loss

---

5. Standardization Logic

5.1 Address Standardization

- Normalize address components into address_line, city, state, and zip_code
- Standardize abbreviations (Street → St, Avenue → Ave)
- Validate state codes and ZIP formats

5.2 Unit and Currency Standardization

- Store all monetary values in USD
- Normalize square footage units
- Preserve original values for traceability where applicable

---

6. Data Fusion Across Sources

When multiple sources provide overlapping information:

- Match records using property_id or standardized address

- Resolve conflicting values using source priority rules or data completeness
- Retain source metadata to support traceability

---

7. Quality Flags and Metadata

Each cleaned record includes quality indicators:

- missing_field_flag
- outlier_flag
- address_validation_flag
- source_conflict_flag

These flags support downstream quality control and analytics.

---

8. Outputs

The cleaning process produces:

- Cleaned and standardized datasets
- Consistent formats across all entities
- Quality flags and metadata
- Data ready for valuation modeling and scoring

---

9. Assumptions and Notes

- Cleaning logic builds directly on normalization rules defined in Task B
- Records are deleted only when required identifiers are missing
- The process prioritizes data consistency and traceability

---

10. Conclusion

The data cleaning and standardization modules ensure ingested real estate data is reliable, consistent, and analytics-ready. This step bridges raw ingestion and downstream valuation and scoring pipelines.