1. Data Scope & Ingestion

- Market Data: Ingests five CSV files from PGE, SCE, and SDGE (Historical & Current) (Source: Interconnected Project Sites Data Set)

2. Exclusions & Filters (The "Zero-Tolerance" Policy)

To ensure the summary math is accurate, the following records are purged before aggregation:

- Status Filter: Only records explicitly marked as "Interconnected" and “Residential” are included.
- Financial/Technical Integrity: Rows with $0 or null Total System Cost and 0 kW or null System Size AC are removed.
- Temporal Integrity: Records with missing or unparseable App Complete Date are discarded.
- ZIP Validation: Strict Regex enforcement (^\d{5}$) removes any ZIP code that is not exactly five digits or contains alphabets/symbols.
- Post-Aggregation Sanity Filter: After aggregation, any **ZIP × month × year** group with market_total_cost < $50 is excluded (to remove rounding artifacts and clearly invalid totals). Rationale is that installed residential PV systems are typically priced on the order of ~$2.5–$3.3 per watt, meaning even a 1 kW system is generally ~$2,500–$3,300 before incentives.

3. Data Harmonization & Transformation

- Numeric Casting: Non-numeric characters (symbols, commas, units) are stripped. Data is cast to float64 to ensure mathematical operations (summing) work correctly.
- Temporal Extraction: Dates are parsed to extract year (YYYY) and month (e.g., Jan, Feb) to facilitate monthly growth charting.
- Geographic Cleaning: ZIP codes are stripped of .0 suffixes and padded with leading zeros to maintain a standard 5-digit string format.

4. Aggregation Logic (The "ZIP-Level Summary")

Instead of raw rows, the data is collapsed. The script groups the data by the following dimensions:

- zip_code, year, month, and status

For every unique group, the following calculations are performed:

- market_total_cost: The mathematical sum of all system costs.
- market_total_kw: The mathematical sum of all system capacities.
- number_of_installations: A count of the number of raw rows that were collapsed into that group.

5. Export Technicals

- Precision Control: Totals are rounded to 2 decimals prior to export (and/or enforced via float_format='%.2f') to keep outputs consistent.