1. Data Acquisition & Scope
- Source: Single master CSV: LowIncome_Applications_Dataset_2026-01-08.csv ([Low Income Solar PV Data](#))
- Time Window (Comparison Dataset): Records are restricted to Apr 2015–Oct 2025 to match the market dataset and enable 1:1 time-series comparison.

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

To maintain high data quality and focus purely on equity-based residential solar, the following filters are applied:
- Status Filter: Only applications with a Current Application Status of "Completed" are processed.
- Financial/Technical Integrity: Any record with $0 or null Total System Cost or 0 kW CEC PTC Rating (KW) is discarded.
- Incentive Sanity Rule: Any record where Incentive Amount exceeds Total System Cost (paid > cost) is excluded.
- ZIP Validation: Strict Regex enforcement (^\d{5}$) removes any entry that is not a clean, 5-digit number, effectively purging alphabetic or international placeholders.

3. Data Harmonization & Transformation
- Numeric Enforcement: String-based currency values and capacity ratings are stripped of symbols ($, ,, kW) and cast to float64.
- Temporal Standardization: The First Completed Date is parsed to generate discrete year and month (short-name) columns for monthly trend analysis.
- Geographic Cleaning: ZIP codes are stripped of extraneous decimal points and leading-zero padded to ensure a consistent 5-character string format.

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

Individual records are collapsed into a structured summary. The data is grouped by:
- zip_code, year, month, and status

For every unique group, the script calculates:
- subsidy_total_cost: Sum of all project costs.
- subsidy_total_kw: Sum of all project capacities (KW).
- subsidy_total_paid: Sum of the actual incentive/subsidy amounts paid out.
- number_of_installations: A count of the raw application rows consolidated into that specific ZIP/Month group.

5. Export Technicals
- Formatting: The export uses float_format='%.2f' to enforce clean 2-decimal precision in the final CSV output.