Cleaning Summary (Section 4.5 Deliverable)

- **Demographics dataset** (demographics_data.csv)
    - **Issue → Action mapping**
        - Aggregate rows for continents / income groups / "World" → removed with _drop_non_country_rows
        - Inconsistent country names and alternative spellings (e.g., "Cape Verde" vs "Cabo Verde") → harmonised; **7** corrections logged to name_mismatches.csv
        - Numeric columns stored as text with thousands-separators → commas stripped and values converted to float via _numericise
        - Missing or implausible life-expectancy values (NaN, < 40 y, > 100 y, negative) → affected rows dropped
        - Duplicate country rows after normalisation → kept first instance, removed the rest
    - **Row count before: 201**
    - **Row count after: 198**
- **GDP per capita PPP dataset** (gdp_per_capita_2021.csv)
    - **Issue → Action mapping**
        - Aggregate rows (continents, income groups, etc.) → removed
        - Country-name mismatches → harmonised using the mapping learned from the demographics step
        - Non-numeric characters in the GDP column (commas, currency symbols) → stripped; values coerced to float
        - Missing GDP values → rows dropped and written to dropped_gdp.csv
        - Statistical outliers identified with Tukey fences (**6** countries) → **flagged but kept**
        - Duplicate country rows → removed
    - **Row count before: 213**
    - **Row count after: 197**

- **Population dataset** (population_2021.csv)
    - **Issue → Action mapping**
        - Aggregate rows → removed
        - Country-name mismatches → harmonised with the same mapping
        - Non-numeric characters in population figures → stripped; values converted to float
        - Missing population values → none detected (no rows dropped)
        - Statistical outliers in $\log_{10}$ scale (**1** country) → **flagged but kept**
        - Duplicate country rows → removed
    - **Row count before: 260**
    - **Row count after: 236**
- **Cross-dataset consistency check**
    - Demographics & GDP = **182 / 198**
    - Demographics & Population = **198 / 198**
    - GDP & Population = **196 / 197**
    - Countries present in **all three** cleaned datasets = **182**