Cleaning summary

Demographics Dataset (`demographics_data.csv`)

- Issue: Invalid life expectancy values (< 40 or > 100)
  → Action: Removed rows outside the valid range
- Issue: Missing life expectancy values
  → Action: Removed rows with missing `life_expectancy_both`
- Issue: Inconsistent country name formatting
  → Action: Normalized using a custom `smart_title()` function with manual exceptions for known special cases
- Issue: Country name mismatches
  → Action: Logged renamed countries to `name_mismatches.csv`

Summary:

- Rows before cleaning: 200
- Rows after cleaning: 200 (no invalid/missing rows found in this case)
- Cleaned data saved to: `output/demographics_data.csv`
- Name mismatches saved to: `output/name_mismatches.csv`

GDP per Capita Dataset (`gdp_per_capita_2021.csv`)

- Issue: Some GDP values may contain commas or non-numeric characters
  → Action: Removed commas and symbols, converted values to numeric using `pd.to_numeric`
- Issue: Missing values (`None`)
  → Action: Removed rows with missing GDP values
  → Dropped rows: 0
- Issue: Potential outliers based on Tukey method
  → Action: Identified outliers using Tukey method (Q1 - 1.5*IQR, Q3 + 1.5*IQR); did not remove them
  → Outliers detected: 6
- Issue: Duplicate country entries
  → Action: Checked for duplicates and retained only the first entry per country
- Issue: Country names inconsistent with demographics dataset

  → Action: Mapped inconsistent country names to match `demographics_data.csv` to avoid loss of data during merging

Summary:

- Rows before cleaning: unchanged (no missing rows dropped) (213)
- Rows after cleaning: same as original minus any duplicates (213)
- Cleaned data saved to: `output/cleaned_gdp.csv`
- Dropped rows saved to: `output/dropped_gdp.csv`

Population Dataset (`population_2021.csv`)

- Issue: Some population values may contain commas or non-numeric characters
  → Action: Removed commas and symbols, converted values to numeric using `pd.to_numeric`
- Issue: Missing values (`None`)
  → Action: Removed rows with missing population values
  → Dropped rows: 0
- Issue: Right-skewed distribution makes raw outlier detection unreliable
  → Action: Applied `log10` transformation and identified outliers using Tukey method on the transformed data
  → Outliers detected: 1
- Issue: Duplicate country entries
  → Action: Checked for duplicates and retained only the first entry per country
- Issue: Country names inconsistent with demographics dataset

  → Action: Mapped inconsistent country names to match `demographics_data.csv` to avoid loss of data during merging

Summary:

- Rows before cleaning: unchanged (no missing rows dropped) (260)
- Rows after cleaning: same as original minus any duplicates(260)
- Cleaned data saved to: `output/cleaned_pop.csv`
- Dropped rows saved to: `output/dropped_pop.csv`