Data Cleaning and Pre-Processing Overview

This document summarizes how the raw datasets—oil production, consumption, proved reserves, and crude prices—were cleaned and transformed into a unified analytical format.

1. Header Detection and Standardization

Raw files had inconsistent header rows. The pipeline automatically located the row containing valid year values and used it as the header, removing all rows above it. Year fields were normalized to clean integers.

2. Cleaning and Standardizing Country Names

Aggregated rows (e.g., World, OECD, Total) were removed. Country names were standardized (e.g., US → United States) and cleaned for whitespace or formatting inconsistencies.

3. Converting Wide Tables into Long Format

All datasets were reshaped from wide (many year columns) to tidy long format: Country | Year | Value. This enabled uniform merging and time-series analysis.

4. Unit Detection and Conversion

Units were inferred heuristically:

- Production and consumption detected as barrels per day if magnitudes were very large, then converted to million barrels per year.

- Reserves detected as million barrels and converted to billion barrels.

5. Cleaning Crude Price Data

The price sheet was inspected for orientation and transposed if necessary. A Brent-like crude price series was extracted, cleaned, and converted into Year | Price_Brent format.

6. Merging Datasets

Cleaned datasets were merged on Country and Year. Diagnostics were generated to identify coverage gaps across variables.

7. Derived Variables

The Reserve-to-Production Ratio (R/P Years) was computed:

R_to_P_Years = (Reserves_Gb × 1000) / Production_Mb

8. Final Quality Checks

Rows with missing keys were removed, numeric values were rounded to four decimals, and coverage flags were added.

The final cleaned dataset is saved as master_oil_panel.csv and is ready for modeling and analysis.