Pipeline Overview

Step 1 — Data Cleaning (raw → cleaned)

The script data_cleaning.py performs the following operations on each dataset:

1. **Timestamp normalization**
    - Conversion to timezone-aware UTC datetime.
    - Chronological sorting and removal of invalid timestamps.
2. **Duplicate removal**
    - Duplicate timestamps are removed, keeping the last observation.
3. **OHLC sanity checks**
    - Rows with missing, zero, negative, or logically inconsistent OHLC values are removed.
4. **Stale quote detection**
    - Long runs of identical close prices (default ≥ 60 consecutive minutes) are identified and removed.
    - This avoids artificially low volatility and distorted feature construction.
5. **FX session filtering**
    - Data are restricted to the FX trading window:
        - **Sunday 22:00 → Friday 22:00 UTC**
    - This step removes non-tradable periods and avoids classifying market closures as data gaps.
6. **Gap detection**
    - Gaps are detected using timestamp differences:
        - **Short gaps**: 1 minute < gap ≤ 2 days
        - **Long gaps**: gap > 2 days (up to 10 days)
    - Gaps are recorded but **not interpolated or filled**, preserving the event-time nature of the data.
7. **Invalid block detection**
    - Consecutive sequences of rows with invalid OHLC values are grouped into blocks and recorded.

The cleaning phase produces:

- cleaned OHLC datasets,
- tables of short gaps,
- tables of long gaps,

- tables of invalid blocks.

Step 2 — Diagnostic Plots (cleaned → figures)

The script plot_diagnostics.py generates three diagnostic figures for each dataset:

1. **Timeline plot**
    - Shows the exact timestamps at which gaps or invalid blocks occur across the full sample.
2. **Duration histogram**
    - Displays the empirical distribution of gap or block durations (in minutes).
3. **Weekday × hour heatmap**
    - Highlights systematic temporal patterns in missing data occurrences.

Figures are saved into subfolders such as:

- Short term_figures/
- Long Term_figures/
- Invalid blocks_figures/

Definitions

- **Short gap**
  A discontinuity between consecutive timestamps where
  1 minute < diff ≤ 2 days.
- **Long gap**
  A discontinuity where
  diff > 2 days, typically corresponding to weekend or holiday closures.
- **Invalid block**
  A contiguous sequence of rows (at 1-minute frequency) where at least one OHLC field is missing or zero.
- **Gap duration**
  Defined as the number of *missing minute bars between two valid observations* (internal missing minutes only).

Key Findings

The diagnostic analysis highlights substantial differences between FX data providers:

- Dukascopy minute-level data exhibit a very large number of short gaps, often corresponding to single missing minutes and distributed across all trading days and hours.
- Long gaps are highly regular in both datasets and closely aligned with weekend closures, with an average duration close to 2,880 minutes.

- IBKR data display a much more stable structure: short gaps follow a deterministic daily rollover pattern, while long gaps are identical across bid and ask series.
- Invalid price blocks are frequent in Dukascopy historical datasets (particularly for AUD/USD and USD/JPY) and are entirely absent in IBKR data. These blocks are artifacts generated by our code used to get the data from Dukascopy
- We confirmed our findings by using live data from Dukascopy(Jforex) and IBKR(TWS). We performed two separated attempts. In the second attempt we got live data for half of a day simultaneously for both data sources, confirming that Dukascopy has many more short-term gaps.
- Most of the short-term gaps in Dukascopy become invalid blocks due to our downloading process. Our code fills every row with the time, but if no data are received from the broker, the row remains empty, generating what we identify as an invalid block.

Example Figures

The resulting graphs are all available in the respective folders. To reduce the size of the entire work, most of the datasets have not been uploaded, while the cleaned datasets are all in “parquet” format for the same reason.