AlpacaTech Take-home Exercise

This assignment contains 2 parts **Data Provider Selection** and **Data ETL**.

Please aim to complete **both** exercises within **an hour**.

In case you exceed an hour in completing the tasks, please indicate the actual time taken in your deliverable.

Data Provider Selection

You are provided with three sample FOREX tick datasets each from a different data provider: A, B and C. They are namely “sample_fx_data_A.csv.gz”, “sample_fx_data_B.csv.gz” and “sample_fx_data_C.csv.gz” in the same folder. Each dataset is provided in a CSV format with the following columns:

- datetime: Date and time of the data point
- currency_pair: Identifier for the currency pair in the format of XXXYYY (e.g., “USDJPY”)
- bid: Bid price of the currency pair at the given timestamp
- ask: Ask price of the currency pair at the given timestamp
- volume: The trading volume during the time interval

Please perform data analysis on the sample datasets, and select the most appropriate data provider with data correctness, data completeness and length in mind.

Deliverables expected:

- **Data investigation report :** A detailed data analysis report in the form of Jupyter notebook (**analysis.ipynb)** and HTML output (**analysis.html)**, with a conclusion indicating and justifying the chosen data provider with reason.

Data ETL

With the chosen data provider's sample data, you are expected to write code that processes the secondly tick data into the **minutely** data with the following attributes:

- timestamp: Date and time of the data point
- currency_pair: Identifier for the currency pair in the format of XXX/YYY (e.g., "USD/JPY")
- open: Opening mid price of the currency pair at the given timestamp
- high: Highest mid price of the currency pair during the time interval
- low: Lowest mid price of the currency pair during the time interval
- close: Closing mid price of the currency pair at the given timestamp

* Note: mid price = (bid + ask) / 2

To cut cost on data storage, we are only concerned with storing USD/JPY and EUR/JPY currency pairs ranging from 2024 onwards. The code is expected to:

1. Read the forex dataset
2. Validate the raw data
3. Clean the data
4. Transform the tick data to minutely OHLC mid price data
5. Validate the processed data
6. Write the results to an output file in a readable format in csv or csv.gz

Deliverables expected:

a) Python Notebook / Program / Script: A python notebook (and HTML output), program or script that contains methods that perform the required ETL process and writes the results to an output file.

b) Output File: A readable output file in CSV format or compressed GZIP format containing the desired output.

c) Readme: Documentation with minimum instructions for running the python program. Optional if the deliverable of (a) is a python notebook. State any assumptions and limitations if necessary.

d) [Optional] Unit Tests: Unit tests to validate the correctness of the code