Quant Evaluation - HFT

注意事项：

- **请不要将该题目与任何人分享！**
- 你一共有2天的时间完成下列问题。你可以**任选一种**方式提交你的结果：
    - （推荐）如果你使用jupyter notebook，请将代码、注释及运行结果一起保存为html格式。
    - 如果你使用其他IDE，请单独提交代码，并将输出结果另外整理成文档，保存为pdf格式。
- 最终评价会从思路、代码、报告形式以及已有经验等多方面综合考虑。
- 在笔试过程中如有任何问题，请及时与我们联系。

Data

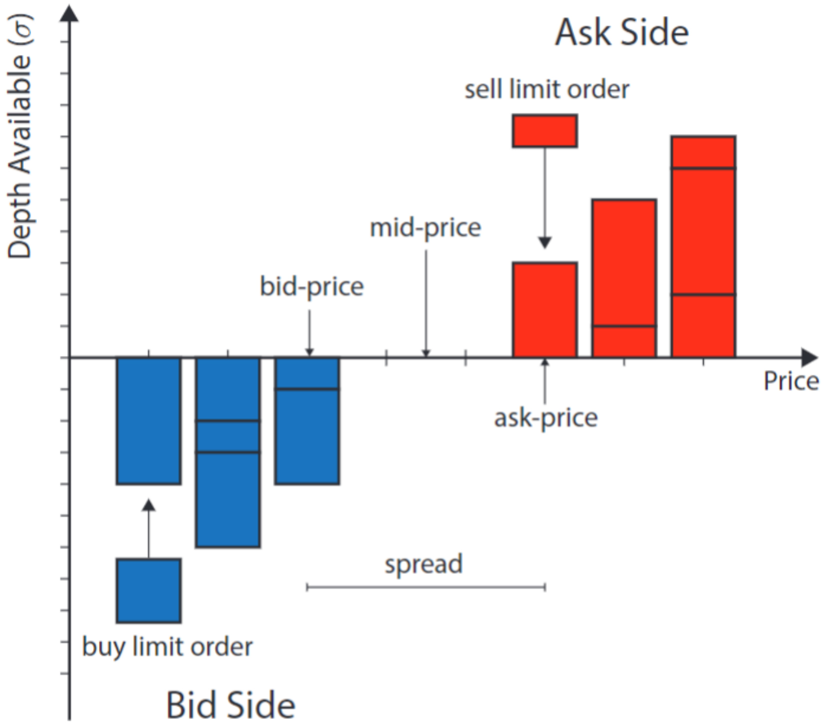
There are 3 `.csv` files in the `./data/` path, which are named as `0.csv`, `1.csv` and `2.csv`. They contain the level-2 snapshots of 3 stocks in 3 consecutive trading days. There are 49 columns in each dataset:

data.zip
1.08MB

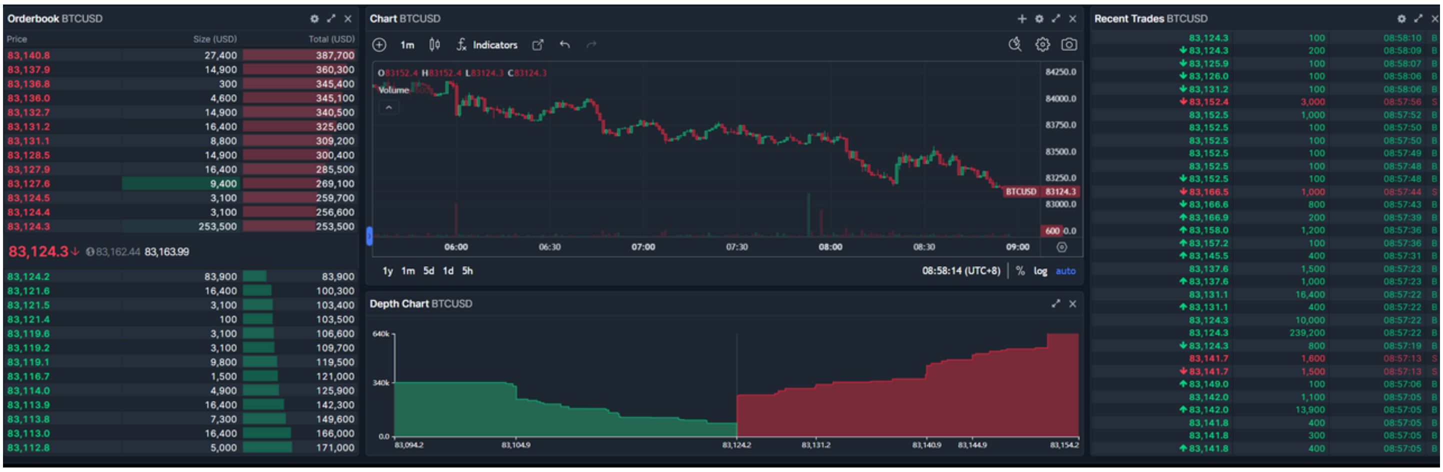
Column Name	Data Type	Meaning
date	int32	
skey	int32	
time	int64	Snapshot time recorded by the exchange. • Unit: millisecond.
volume	int64	Cumulative trading volume: The total trading volume that has occurred up to the current snapshot.
amount	float64	Cumulative trading value: The total trading value that has occurred up to the current snapshot.
open/high/low	float32	Open/high/low price.
last	float32	The latest trade price.

bp1/bp2/.../bp10	float32	Top 10 bid prices in the order book.
ap1/ap2/.../ap10	float32	Top 10 ask prices in the order book.
bq1/bq2/.../bq10	int32	Top 10 bid quantities in the order book.
aq1/aq2/.../aq10	int32	Top 10 ask quantities in the order book.

An **order book** is a list of orders that a trading venue uses to record the interest of buyers and sellers in a particular financial instrument. Whenever an institution submits a buy (sell) order x , a LOB's trade-matching algorithm checks whether it is possible for x to match a sell (buy) order y such that $p_y \leq p_x$ ($p_y \geq p_x$). If so, the match occurs immediately and the owners of the relevant orders reach a trade for specified amount at specified price. Otherwise x is inserted at price p_x , until it either matches an incoming sell (buy) order or is cancelled by its owner. The figure below shows a slice of the order book:



You can also visit some exchange websites to take a look at charts for real-time order books, which might help you better understand it.



Orderbook BTCUSD

Price	Size (USD)	Total (USD)
83,140.8	27,400	387,700
83,137.9	14,900	360,300
83,136.8	300	345,400
83,136.0	4,600	345,100
83,132.7	14,900	340,500
83,131.2	16,400	325,600
83,131.1	8,800	309,200
83,128.5	14,900	300,400
83,127.9	16,400	285,500
83,127.6	9,400	269,100
83,124.5	3,100	259,700
83,124.4	3,100	256,600
83,124.3	253,500	253,500
83,124.3 ↓	83,162.44	83,163.99
83,124.2	83,900	83,900
83,121.6	16,400	100,300
83,121.5	3,100	103,400
83,121.4	100	103,500
83,119.6	3,100	106,600
83,119.2	3,100	109,700
83,119.1	9,800	119,500
83,116.7	1,500	121,000
83,114.0	4,900	125,900
83,113.9	16,400	142,300
83,113.8	7,300	149,600
83,113.0	16,400	166,000
83,112.8	5,000	171,000

Chart BTCUSD

Recent Trades BTCUSD

83,124.3	100	08:58:10	B
↓83,124.3	200	08:58:09	B
↓83,125.9	100	08:58:07	B
↓83,126.0	100	08:58:06	B
↓83,131.2	100	08:58:06	B
↓83,152.4	3,000	08:57:56	S
83,152.5	1,000	08:57:52	B
83,152.5	100	08:57:50	B
83,152.5	100	08:57:50	B
83,152.5	100	08:57:49	B
83,152.5	100	08:57:48	B
↓83,152.5	100	08:57:48	B
↓83,166.5	1,000	08:57:44	S
↓83,166.6	800	08:57:43	B
↑83,166.9	200	08:57:39	B
↑83,158.0	1,200	08:57:36	B
↑83,157.2	100	08:57:36	B
↑83,145.5	400	08:57:31	B
83,137.6	1,500	08:57:23	B
↑83,137.6	1,000	08:57:23	B
83,131.1	16,400	08:57:22	B
↑83,131.1	400	08:57:22	B
83,124.3	10,000	08:57:22	B
83,124.3	239,200	08:57:22	B
↓83,124.3	800	08:57:19	B
83,141.7	1,600	08:57:13	S
↓83,141.7	1,500	08:57:13	S
↑83,149.0	100	08:57:06	B
83,142.0	1,100	08:57:05	B
↑83,142.0	13,900	08:57:05	B
83,141.8	400	08:57:05	B
83,141.8	300	08:57:05	B
↑83,141.8	400	08:57:05	B

To simplify the analysis, only data during the consecutive trading period (09:30:00~11:30:00 and 13:00:00~14:56:57) are provided. **You MUST analyze all 3-day data** to enhance the credibility of your conclusions.

Problems

1. Perform exploratory data analysis on the dataset.
2. Calculate following statistics of each stock. Please print your results.
   a. 3-day highest/lowest price price.
   b. 3-day cumulative return/trading amount.
   c. Total number of BBO price changes in 3 days. Here BBO price change is defined as `bp1` or `ap1` changes.
3. Ideally, we can receive level-2 snapshot of each stock every 3 seconds. However, in reality, some snapshots may be missed. Please design a suitable data cleaning method so that the new dataset starts at 09:30:00 and ends at 14:56:57, the frequency of which is 3 seconds. **The new dataset for each stock on each day should contain 4,741 rows.** We treat it as a standard dataset.

Hint: Please use the standard level-2 snapshot data to answer the following questions.

4. Let's define

   `mid_price = (bp1 + ap1) / 2`,

   `spread = (ap1 - bp1) / mid_price`.

   Please answer the following questions

   a. Show the distribution of `spread` data. Report 1%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, 99% percentiles.
   b. Summarize meaningful phenomena, such as

    i. Trends among different days.

    ii. Trends among different stocks.

    iii. Reasons for sudden increases in certain snapshot.

5. Choose your optimal solution to define the volatility of the past 20 snapshots for any given time t. (**Hint**: You should take the difference between fluctuating increasing/decreasing and rapid increasing/decreasing into consideration, and analyze the pros and cons of your definition.)
6. Let's define

   `trading volume between 2 adjacent snapshots = active buy volume + active sell volume`

   `active buy = trades of which price ≥ ap1`

   `active sell = trades of which price ≤ bp1`

   Please choose your optimal solution to estimate the active buy and sell volume between 2 adjacent snapshots, and list possible issues.
7. Define cancellation as the number of pending orders cancelled at each price level on the order book between 2 adjacent snapshots. According to the **10-level** order book and the estimated active buy and sell volume, estimate the cancellation quantity **at each price level** between 2 adjacent snapshots.
8. `mid_price` is the simplest way to define the price center of a snapshot. Can you think of a better solution to define the price center `weighted_price`? Please explain the logic behind it and provide examples to illustrate the advantages of `weighted_price` over `mid_price`.
9. Let's define

   `the imbalance of buy and sell in the order book = (bq1 - aq1) / (bq1 + aq1)`.

   Logically, if more investors want to buy the stock, the value of this factor will be larger, and vice versa. We believe that this imbalance factor is highly correlated with the future return. Can you think of a better factor to measure the imbalance of buy and sell in the order book? Please explain the logic behind it and analyze the relationship between imbalance and future return **through statistical tests**.
10. Up to now, you should have gained some knowledge of high-frequency trading. Please write a summary of all your takeaways. It will also be appreciated if you want to share your comments on this written test, or anything else like possible profitable strategies. Hope to meet you in the next interview.