Note that you are forbidden from distributing or referencing this document or any of its attachments to any third parties.

Introduction:
Please address the questions below in a clear and concise write-up. Feel free to use tables and graphs to illustrate your findings. Please note that though you should attach your code, the write-up remains your most important presentation.

Data:
mdLog.csv contains *snapshots* of market data from a single security. Note that this means you are not provided with new information each time the market changes. Instead, you are given periodic and regular updates. The allocation mechanism of this market follows *price-time priority*, incoming orders must interact first with the opposing order with the best price, followed by the earliest order insertion time.

- updateCount: a consecutive counter.
- lastprice: the execution price of the most recent trade in the market from the perspective of the current update.
- volume: the cumulative volume traded in the security from the beginning of the dataset to the current update.
- bid/ask: the price at the best bid / best ask.
- bidsize/asksize: size at the bid/ask.

Questions:

1. How many times did the bid-ask spread (defined as the ask price less the bid price) widen in this dataset? What proportion of those times did it widen on both sides (bid decreased **and** ask increased)? What proportion of those times did it widen on one side (bid decreased **xor** ask increased)? How many times did the bid-ask spread tighten? We define market size as (bidsize + asksize)/2. Report the distribution of market size.

2. Focus on instances in which the mid price changes but the bid-ask spread does not. For an increase (decrease) of the mid price, define the new bid (ask) as the aggressive side, and the other as the defensive side. Report the distribution of the market sizes of the defensive side and aggressive side immediately following a change in mid price. Can you intuitively explain your results?

3. If, from one update to the next, volume increases from 100 to 150, we know 50 was traded between the two updates. Many trades of varying size at various prices could have happened between the two updates; we only know that the most recent trade happened at the lastprice. Come up with a *realistic* volume allocation method that allocates the traded volume to price levels. Justify the method and append your allocation to mdLog.csv.

4. We define sizeDelta as the size added or cancelled on a level from one update to the next, net of any traded volume. For example, on update 1, the bidsize was 100. On update 2, the bid price is unchanged and the bidsize is now 70. Suppose, from part 3, that we believe a total of size 20 was traded at the bid. Then we have sizeDeltaAtBid = -10, as in orders totaling size 10 were canceled between update 1 and update 2 on the bid. Note that sizeDelta is only defined for price levels that are unchanged from the previous update.

Add to mdLog.csv two additional columns: sizeDeltaAtBid and sizeDeltaAtAsk. Report the distribution of these two variables.

5. Now we will focus on the aggressive side. Assume the size on the aggressive side we see immediately after a price change is from a single limit order, which we call the top order. We wish to track the performance of top orders. Assume cancellations always happen from the back of the queue and the order may be partially cancelled.

For example:

update	Bid price	bidsize	Volume allocated to 20.05	sizeDeltaAtBid	comment
1	20.05	50	-	50	Assume this is the aggressive side, so the top order is of size 50
2	20.05	100	10	60	the top order was filled 10
3	20.05	20	10	-70	the top order was filled another 10 for a total of 20 traded, and there must have been 10 cancelled from the top order (because the bidsize is now 20)
4	20.05	40	20	40	20 more was filled at the bid. The rest of the top order is now completely filled.

Note that all top orders are inserted at the BBO (Best Bid Offer), as in a buy (sell) top order is always inserted at the best bid (ask). However, not all top orders spend their entire lifespan on the BBO. Report the total number of top orders, the number that do not spend their entire lifespan on the BBO, and the number that do. Out of those that do spend their entire lifespan on the BBO, how many are filled for their original size?

6. We define return as the signed difference between the execution price of an order and the mid price 40 market updates after the time of the final trade. By convention, we normalize these returns by the ticksize, which is 0.01 in this case. Therefore, if a buy order is all traded at 19.20, the bid 40 updates later is 19.20, and the ask 40 updates later is 19.21, then the return is 0.5. Report average returns of top orders that spend their entire lives at the BBO and are completely filled.

At this point, you may tackle none, some or all of the following questions. You may answer qualitatively or quantitatively.

7. Shift focus to top orders that do not spend their entire lives at the BBO. Report the average returns of these orders. Explicitly state your assumptions regarding your treatment of these orders while they are not at the BBO.

8. Focusing on top orders that spend their entire lifespan at the BBO and are completely filled, can you write a model for the return of these orders based on the size of such orders and the sizeDelta variables you see on the second update after the order was inserted? Can you improve your model via other predictors with information taken on or before the second update after the order was inserted?

9. What are the pros and cons of using the above definition of returns as a metric for order attractiveness? What are some (potentially better) alternatives?

10. Any other interesting observations you want to point out? What assumptions in the problems struck you as unrealistic or overly simplistic?