SCIENTECH

Instruction to download the data

You can download the data from the links below:

https://scientech-pub-training.s3.amazonaws.com/quant-test/trade.csv.gz

https://scientech-pub-training.s3.amazonaws.com/quant-test/quote.csv.gz

Description of the data

The data is a sample of market data (quote and trade) from Shenzhen Stock Exchange on 2020-10-09. This data is collected in real-time and is not guaranteed to be perfect.

The following columns are given in the quote file (quote.csv.gz):

- recv_time: the timestamp in epoch time.
- symbol: the name of the stock.
- bid_price: the best bid price, i.e. the highest price that a buyer would pay for a stock.
- bid_size: the quantity (in share) available at the best bid price.
- ask_price: the best ask price, i.e. the lowest price that a seller would accept for a stock.
- ask_size: the quantity (in share) available at the best ask price.

The following columns are given in the trade file (trade.csv.gz):

- recv_time: the timestamp in epoch time.
- symbol: the name of the stock.
- trade_price: the price of the trade.
- trade_qty: the quantity (in share) of the trade.

Questions

1. Perform exploratory data analysis on the dataset.

2. Find the stock with the largest trade volume between 09:30 to 10:00 (China Standard Time). Show the corresponding largest volume as well.

3. Find the stock with the largest trade volume within each 30-minute window from 09:30 to 15:00, i.e., 09:30 to 10:00, …, 14:30 to 15:00 (China Standard Time). Show the corresponding largest volume as well.

4. Find the top 5 stocks with the largest total trade volume. Plot the series of per-minute trade volume for each of these 5 stocks.

5. Compute the mean spread for symbol 000021.SZ on quotes. Here spread(t) = ask_price(t) - bid_price(t), where t is the timestamp.

6. Compute the median spread for symbol 000021.SZ on trades. Here spread(t) = ask_price(t) - bid_price(t), where t is the timestamp.
Hint: you need to merge quotes and trades to find out bid_price and ask_price of the same symbol at the time when the trade happens.

7. Compute the t-stat of 5-minute open-to-open log-returns for each stock on quotes. The null hypothesis is that the mean of those log-returns is 0. Several steps are involved to solve this problem:

a. Group the quotes into multiple 5-minute windows.

b. Obtain the series of open_mid_price (i.e. the first mid_price you observe in each window), where mid_price(t) = 0.5 * (bid_price(t) + ask_price(t))).

c. Calculate the t-stat of 5-minute open-to-open log-returns, where log-return(this_window) = log(open_mid_price(next_window) / open_mid_price(this_window)).

8. Train a model that best predicts the 10-second log-returns of mid_price on quotes and trades. Here mid_price(t) = 0.5 * (bid_price(t) + ask_price(t)) and 10-second log-return(t) = log(mid_price(t+10 seconds) / mid_price(t)). Please split the dataset into a 60%-40% partition for training and test sets. Describe the fitted model (target variable, features, summary statistics, model choice and hyper-parameters tuning, if any), present both training and test performance, and compare different approaches if you try multiple models.

Hint: you might want to consider shifting price and using merge_asof to compute log-returns.

Non-disclosure Agreement (NDA)

Confidential Information

The content of the coding test ("Test"), including questions, answers or any communication, regarding or related to the Test is Scientech Research’s confidential information ("Confidential Information"). You are expressly prohibited from disclosing, publishing, posting or selling any Confidential Information, in whole or in part, in any form or by any means, oral or written, electronic or mechanical, for any purpose. The submission of the Test expresses and constitutes your entire understanding with respect to the subject matter of this Non-disclosure Agreement.