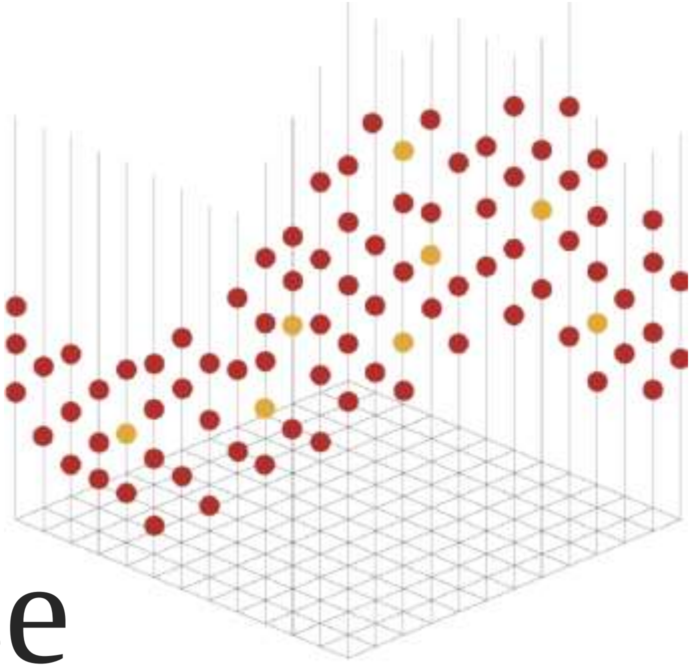
Data Science Institute



Trading At the Close

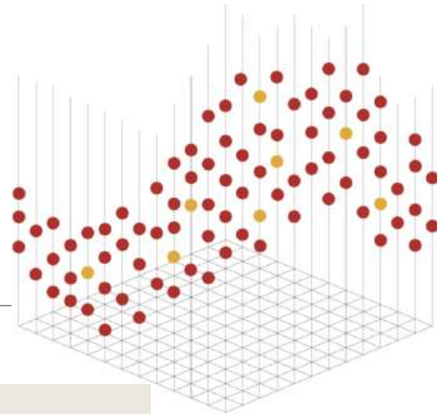
-- Predict US stocks closing movements

DATA1030 MIDTERM PRESENTATION: YU, LETIAN

BROWN UNIVERSITY

GitHub: https://github.com/LetianY/data1030-optiver-trading-at-close/

Introduction



NASDAQ Stock Market:

- Rapid price change in last 10 min
  (10% of average daily volume!)

- **Dataset:** historic data for the daily ten minute closing auction

- **Data Source:** Kaggle by Optiver

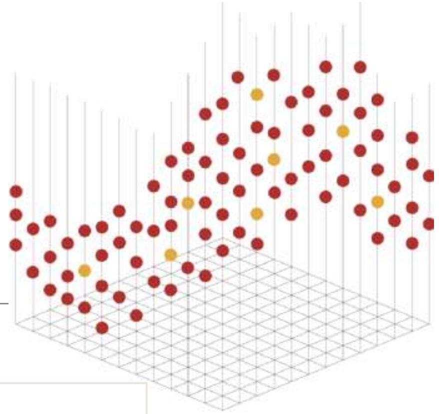
- **Data Collection:** order books and the closing auctions of the stocks

- **Goal:** predict closing price movements for hundreds of listed stocks

- **Problem Type:** Regression

- **Target:** synthetic index
  (closing price movement)

- **Importance:**
  - prices adjustment
  - supply and demand dynamics
  - trading opportunities



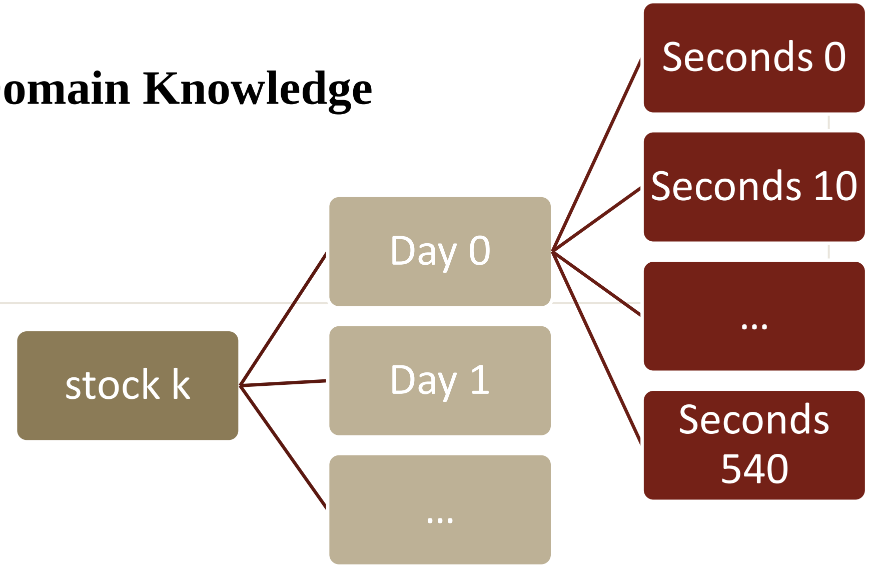
Challenges

- **Missing data:** time structure & features

- **Time series data:** non-iid

- **Large dataset:** 5M+ data points

- **Domain Knowledge**



Data shape: 5,237,980 * 17

- 1 target variable
- 5 identifiers (stock & time)
- 11 market features

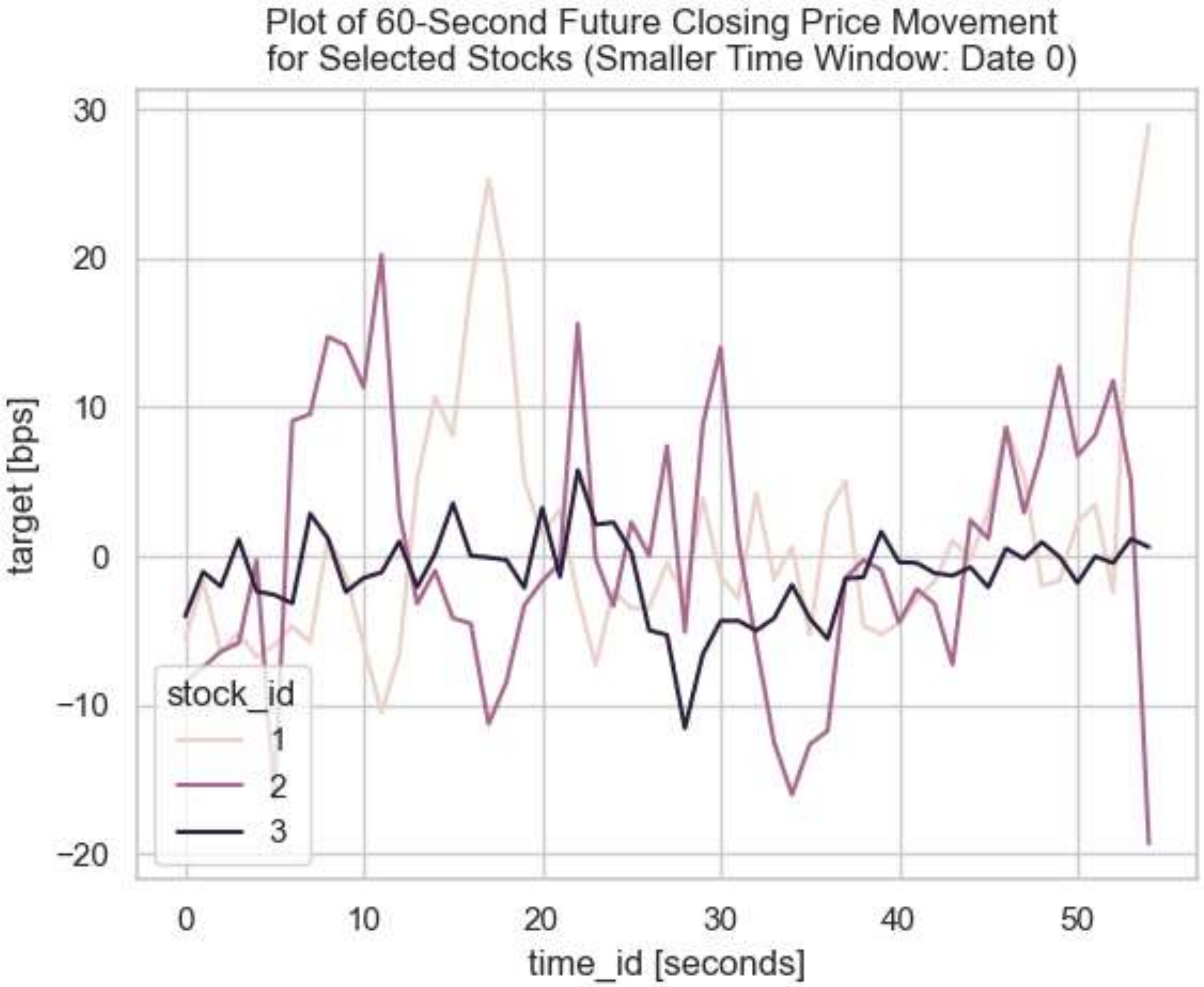
Non-auction book: e.g., bid/ask price
Auction book: imbalance size, reference price, matched size, far price
Auction + non-auction book: near price

Exploratory Data Analysis I

- Volatility

- Extreme Values

- Mean Reversion

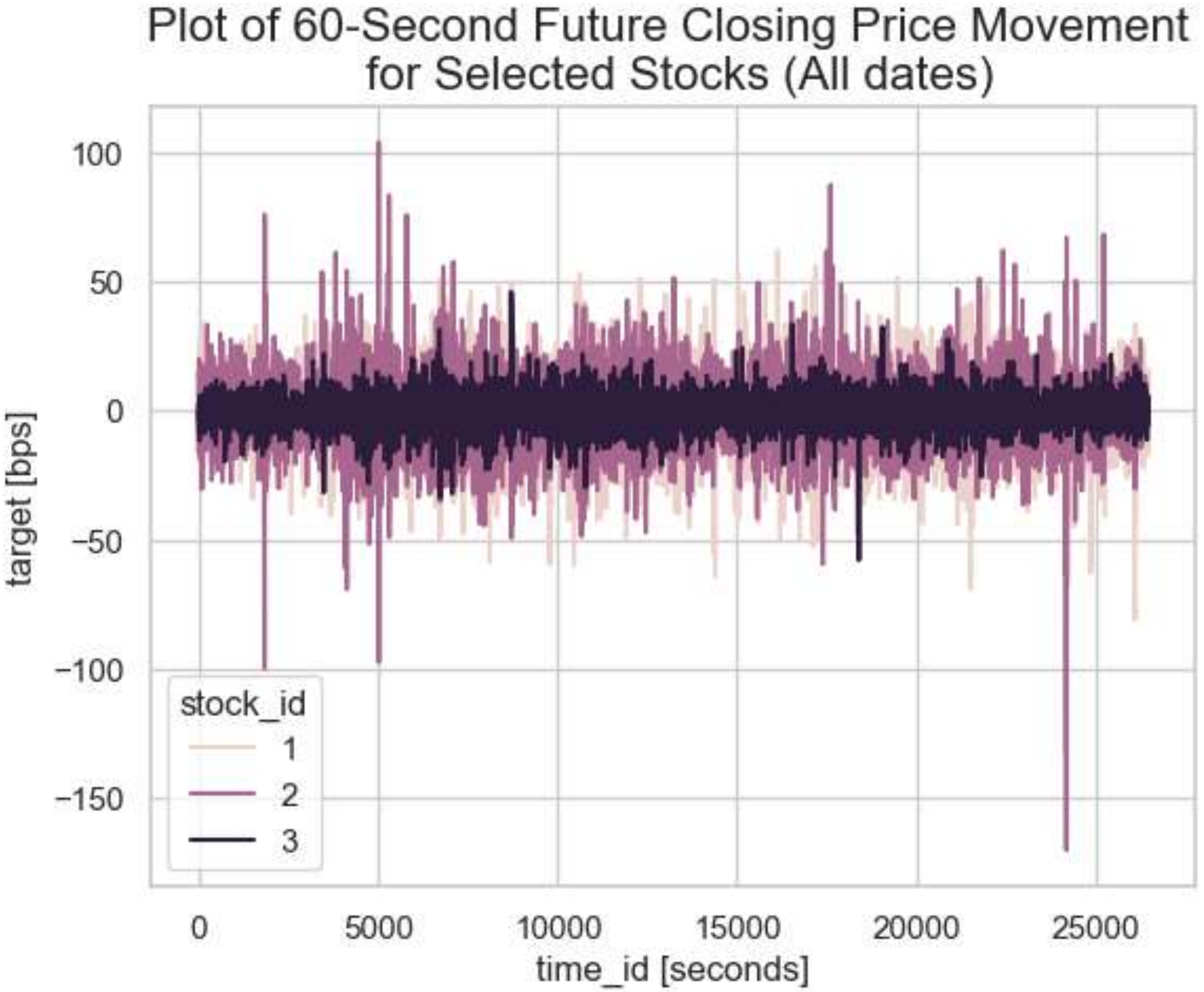


Letian Yu
Brown DSI

Exploratory Data Analysis I

- Volatility
- Extreme Values
- Mean Reversion

→ Test autocorrelation later!



Plot of 60-Second Future Closing Price Movement for Selected Stocks (All dates)

The plot shows the time series plot of the target 60-second future closing price movement index for selected stocks. We see that different stocks shows different volatilities and there exists extreme values. But in general, mean reversion towards zero is perceived.

Exploratory Data Analysis II

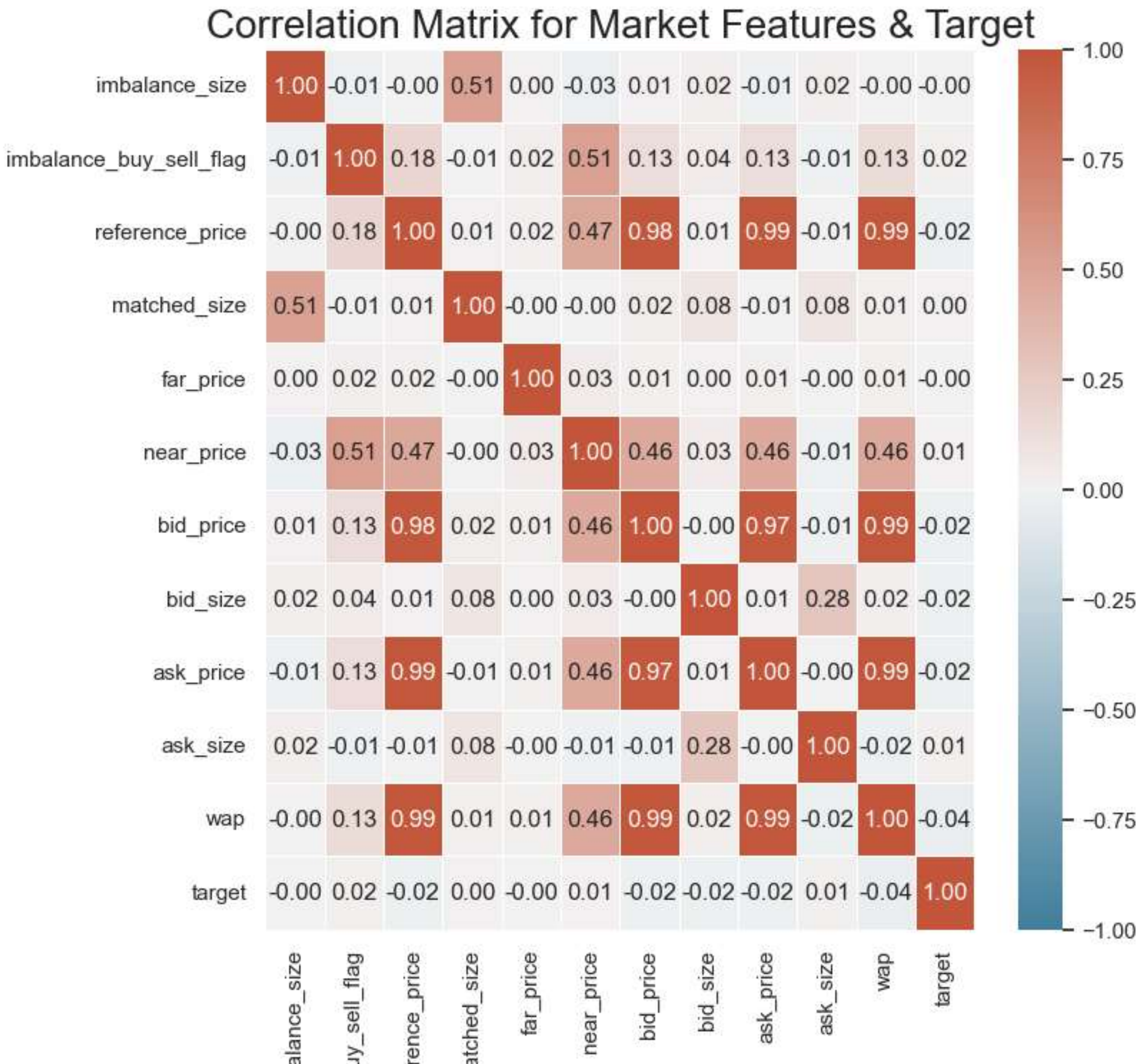
Strong correlation between:

- Bid price, ask price, reference price, wap

- These price are closely related in definition!

- They are also converted to a relevant price

→ Test whether to remove features!



Exploratory Data Analysis II

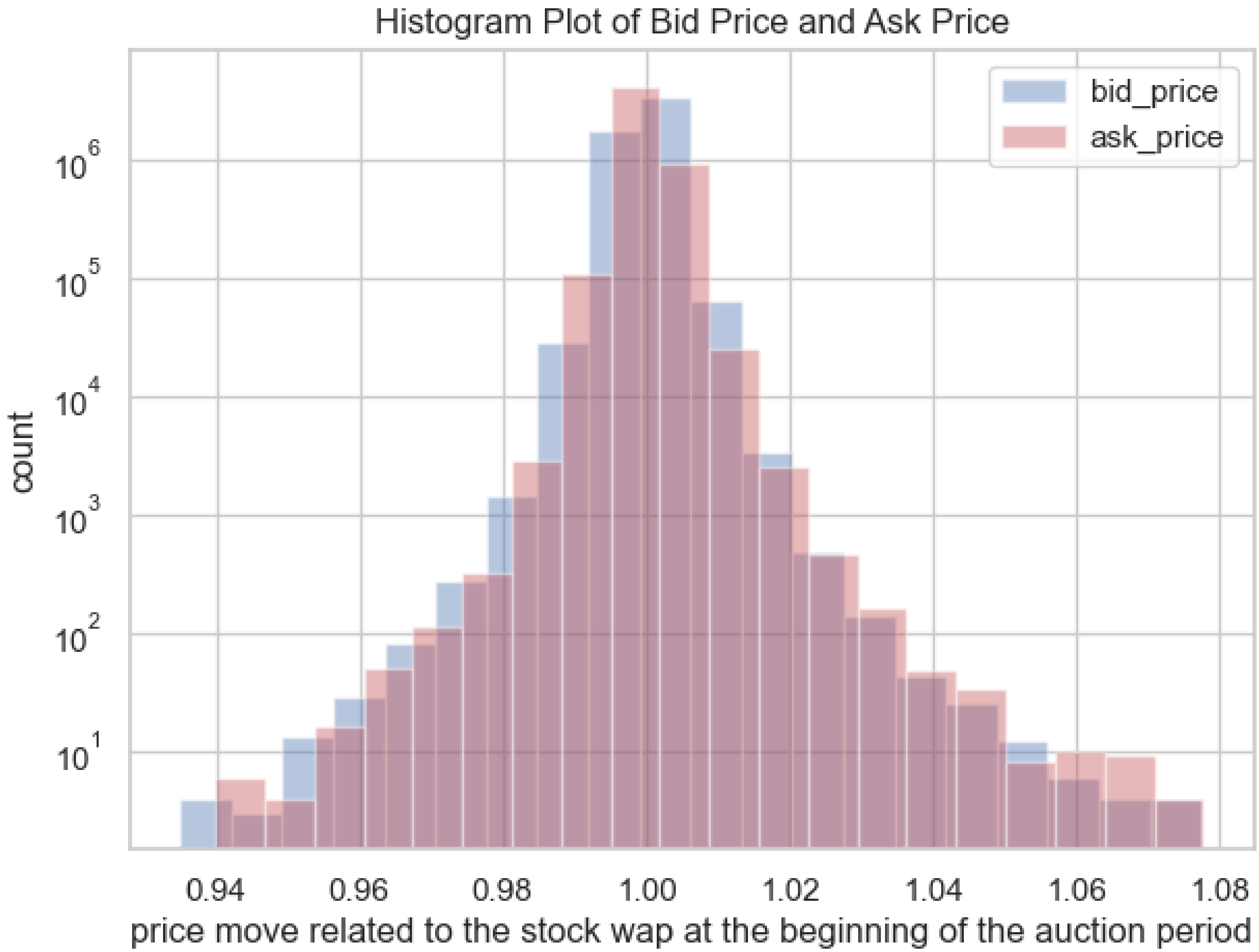
Strong correlation between:

- Bid price, ask price, reference price, wap

- These price are closely related in definition!

- They are also converted to a relevant price

→ Test whether to remove features!



Letian Yu
Brown DSI

Exploratory Data Analysis II

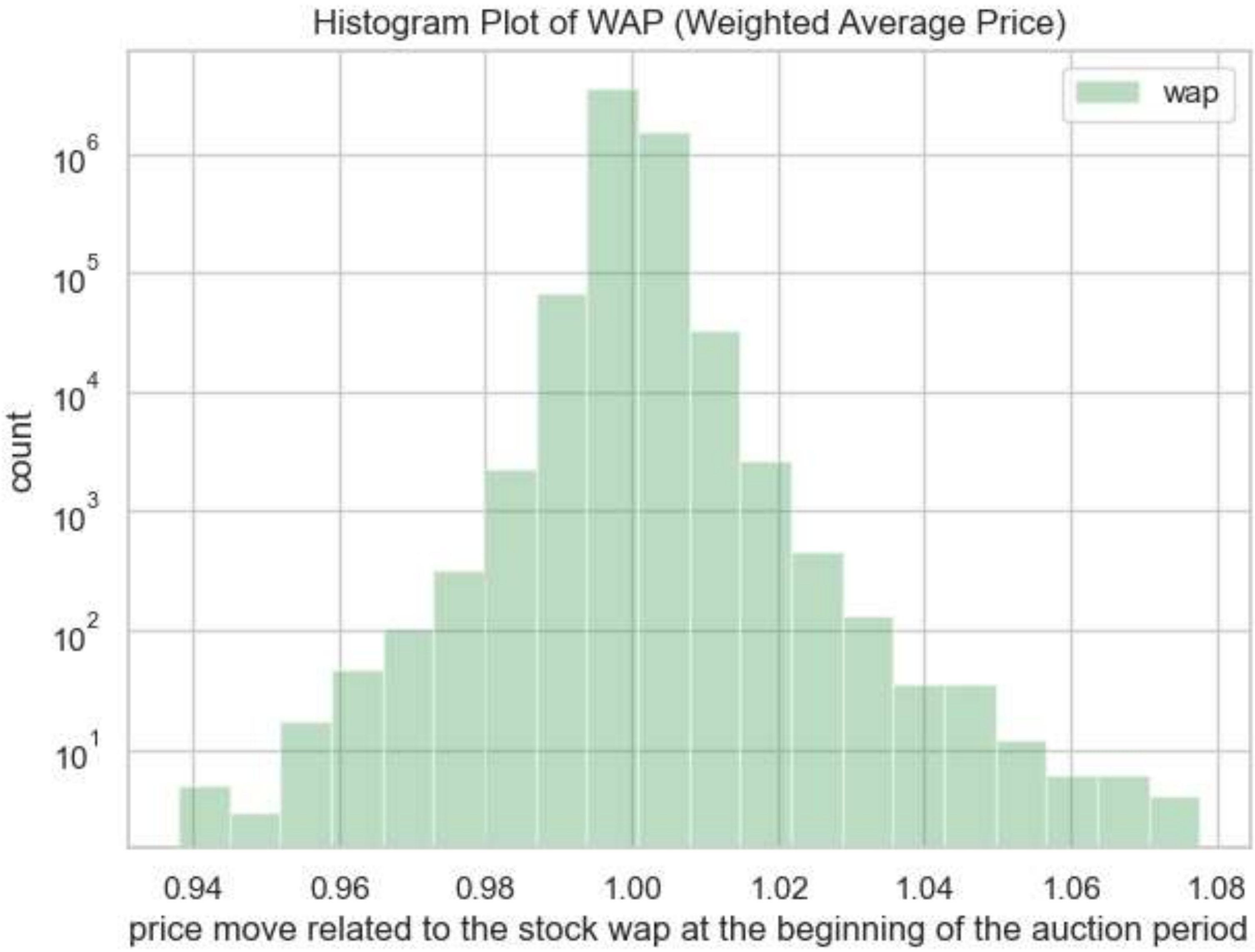
Strong correlation between:

- Bid price, ask price, reference price, wap

- These price are closely related in definition!

- They are also converted to a relevant price

→ Test whether to remove features!



Letian Yu
Brown DSI

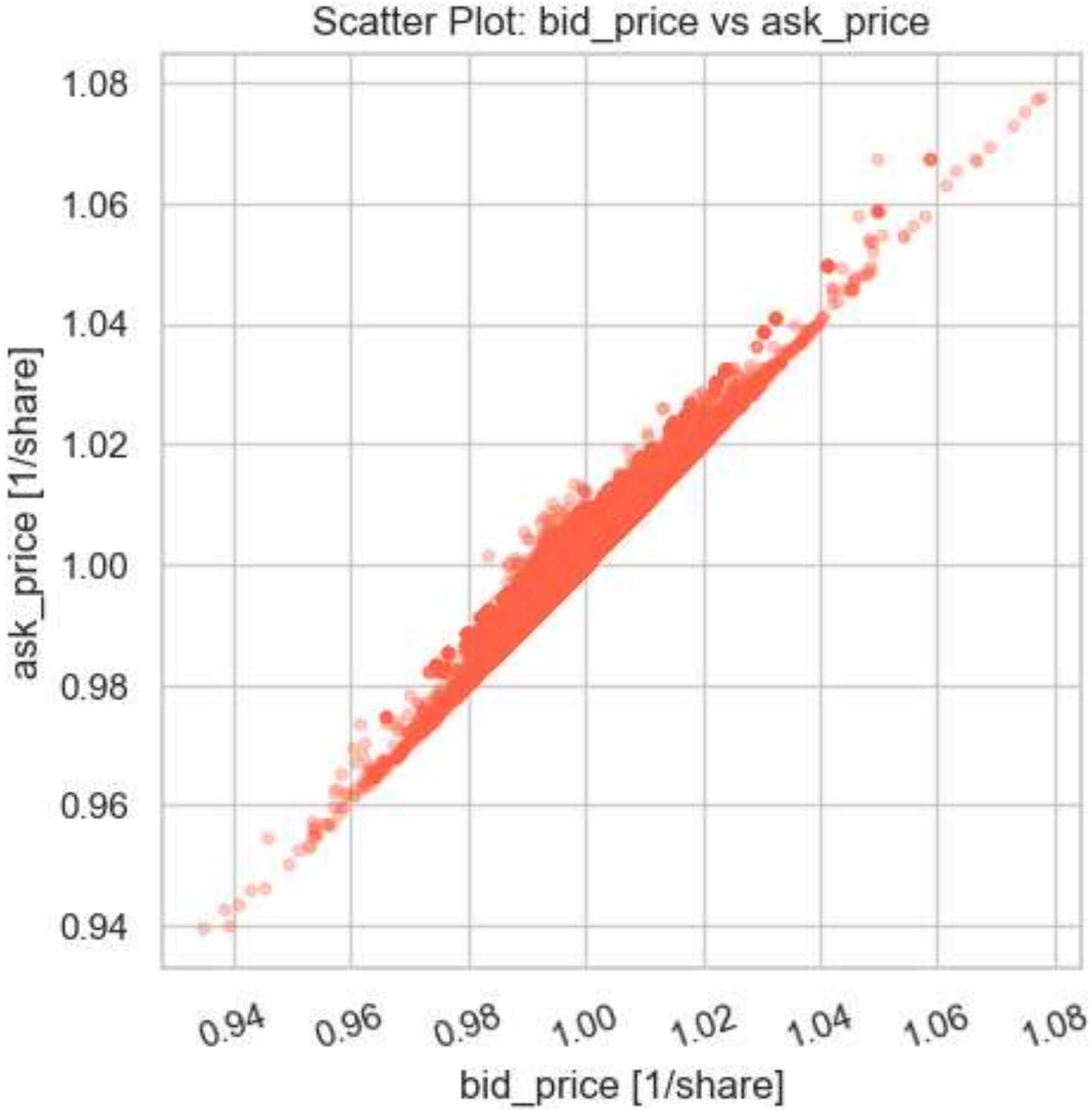
Exploratory Data Analysis II

Strong correlation between:

- Bid price, ask price, reference price, wap

→ In non-auction book, bid price is always smaller than ask price!

→ We may construct extra feature from this!



Note: here both bid price and ask price are a converted price move related to stock wap at the beginning of the auction period. We now see it's always the case that bid price <= ask price.

Data Splitting

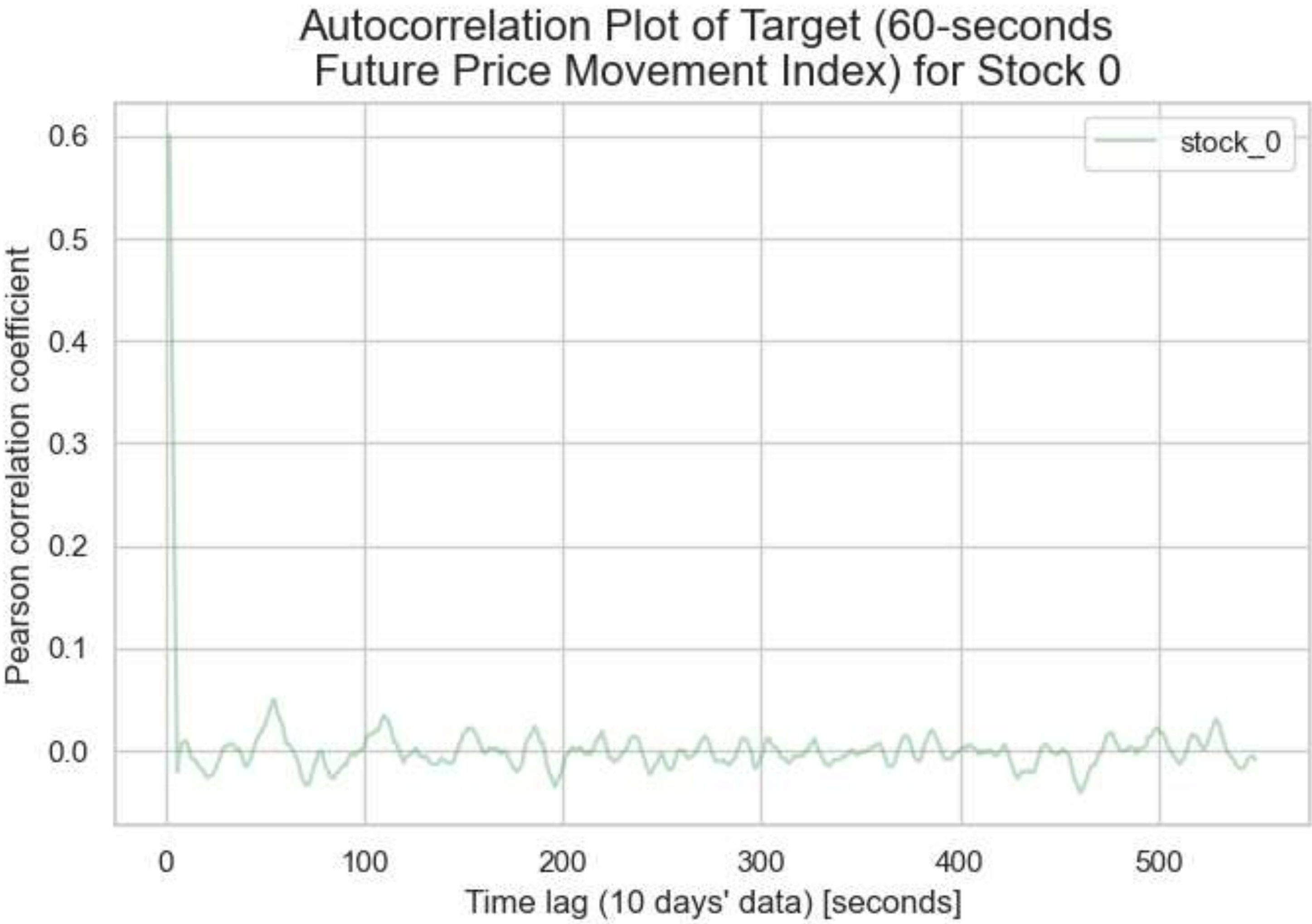
Autoregression & Lagged Features:

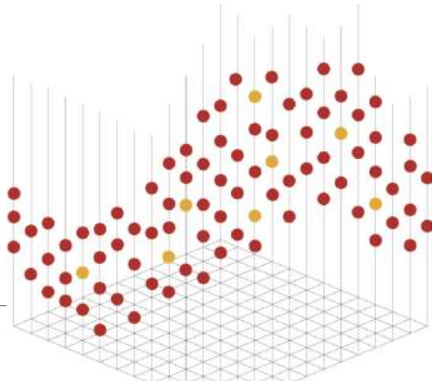
- Multi-stock, have other features

- I followed the real-world and competition setting

- Only previous-day target data is available!
  - we don’t know how the synthetic index is generated
  - Avoid data leakage

- 55 lagged features in total





Data Splitting

GroupTimeSeriesSplit

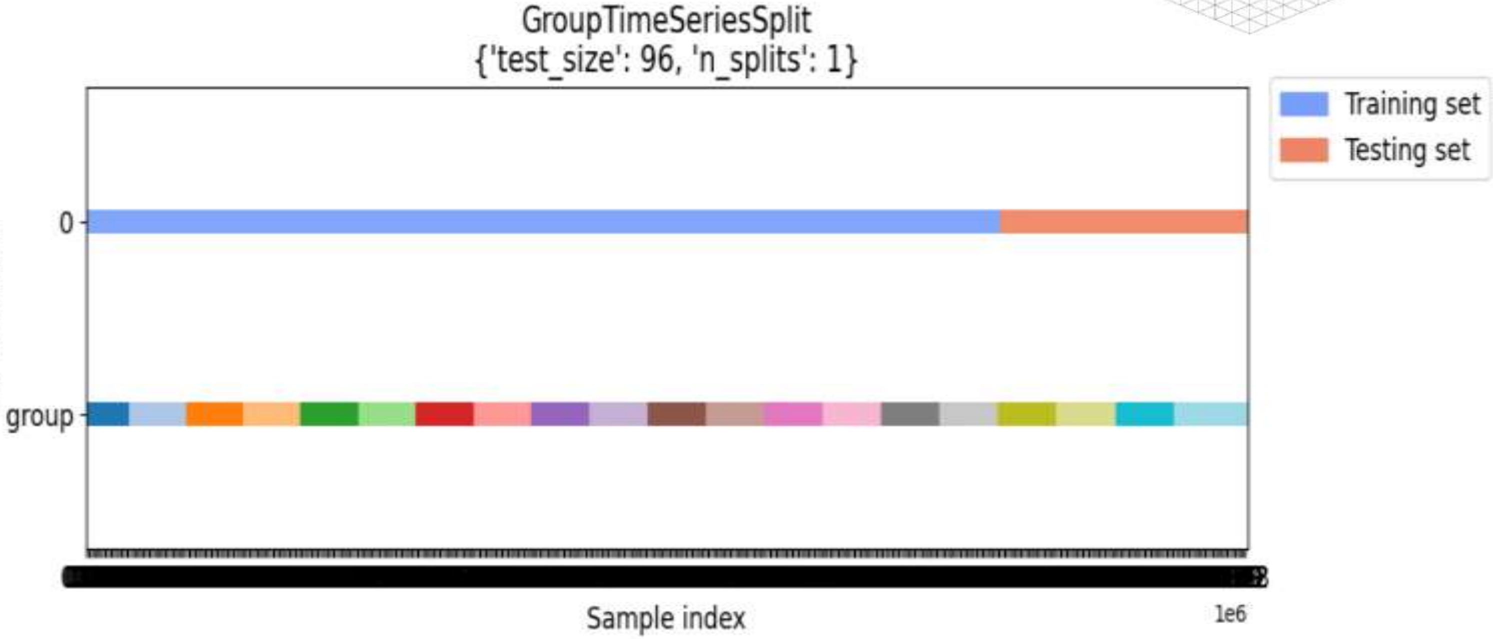
- ML Extension for sklearn

- date_id chosen as group

Why?

- Group and Time Series Structure of Data

- Follow competition setting



Preprocessing

Missing Values for Training Set:

Target missing count: 32 - 0.0007% of data

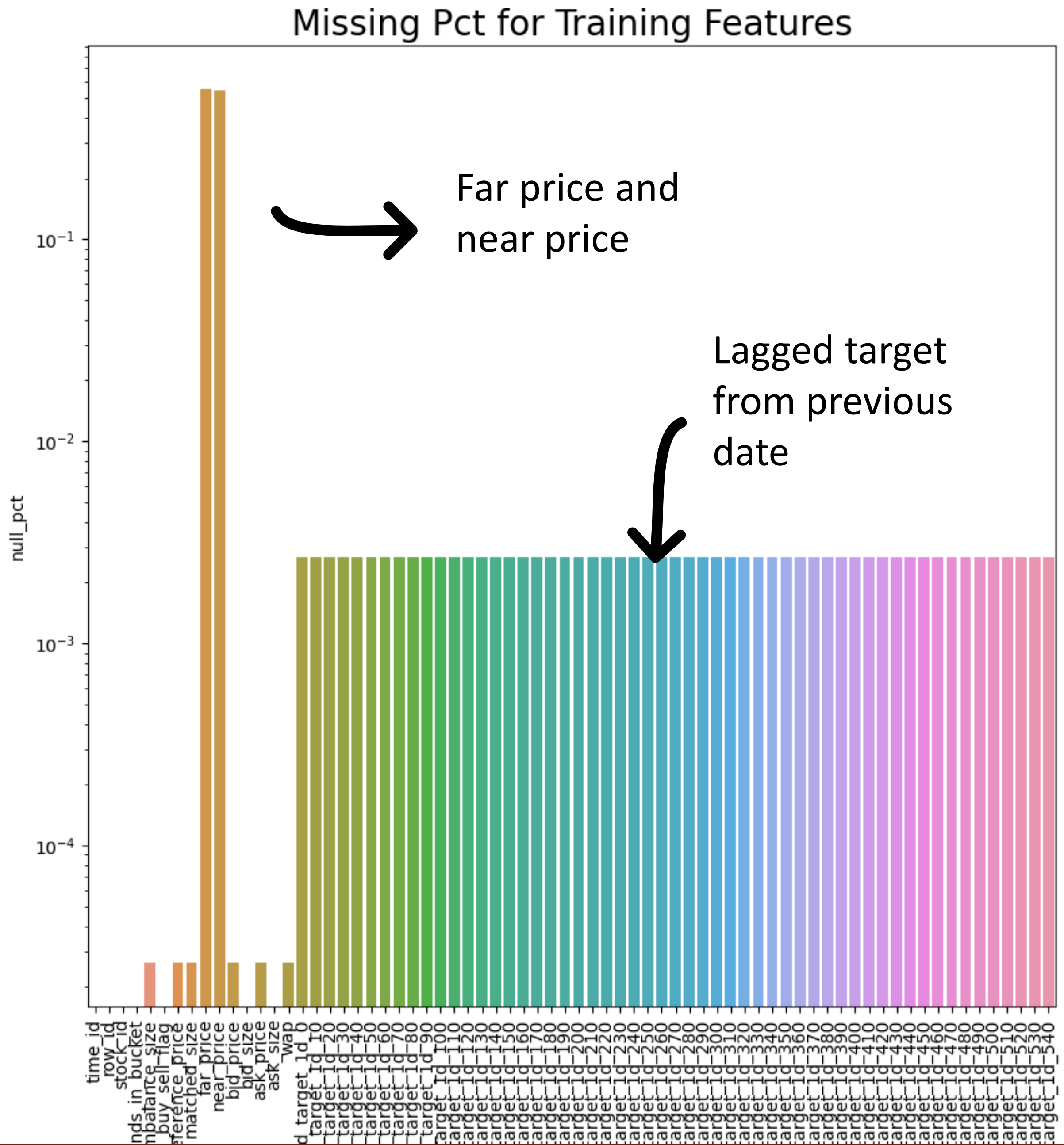
Far price & near price: about 55% missing

Ask price, imbalance size, reference price, matched size, bid price, wap: we have 110 missing, less than 0.001% of total points

Lagged columns: 0.027% missing

Columns: 64 out of 72 columns have missing

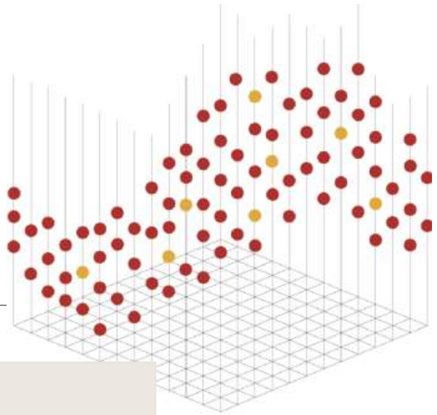
Rows: over 55 percent of data have missing



Far price and near price

Lagged target from previous date

Preprocessing



One Hot Encoder for Categorical Features:

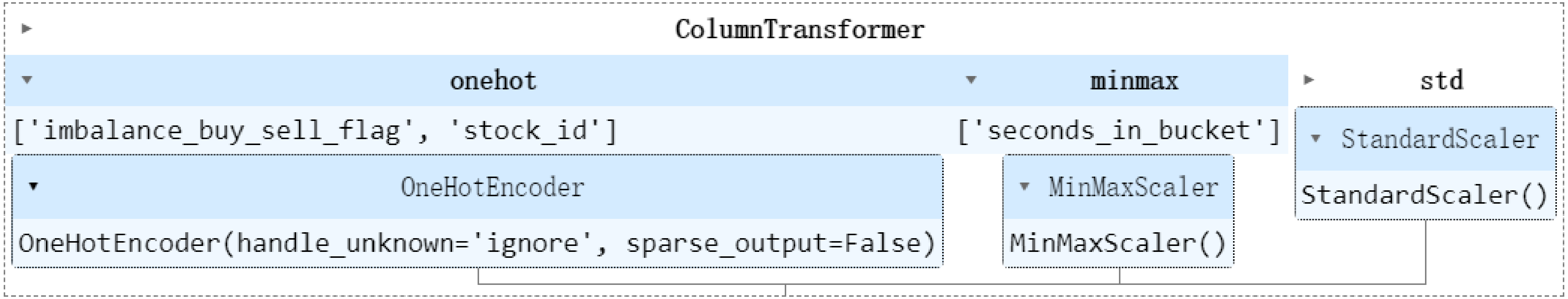
- stock id, buy and sell imbalance flag

MinMax Scaler for time:

- seconds in bucket (originally every 10 from 0 to 540)

StandardScaler:

- Other continuous features:



```
ColumnTransformer
onehot                                                  minmax                    std
['imbalance_buy_sell_flag', 'stock_id']                 ['seconds_in_bucket']     StandardScaler
OneHotEncoder                                           MinMaxScaler              StandardScaler()
OneHotEncoder(handle_unknown='ignore', sparse_output=False)   MinMaxScaler()
```

1. Problem Statement & Data Description

2. EDA:

- Volatility and Mean Reversion of Target
- Close Relationships Between Bid, Ask, WAP price

3. Splitting:

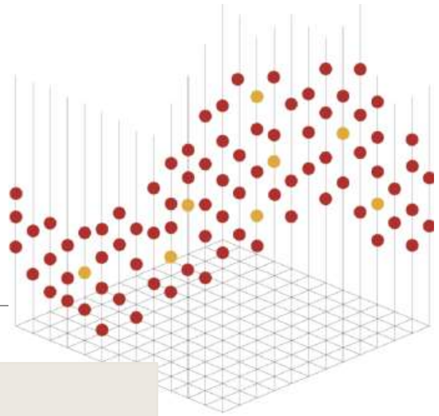
- Autoregression & Lagged Features
- GroupTimeSeriesSplit

4. Preprocessing:

- Missing Values
- Preprocessors

Summary

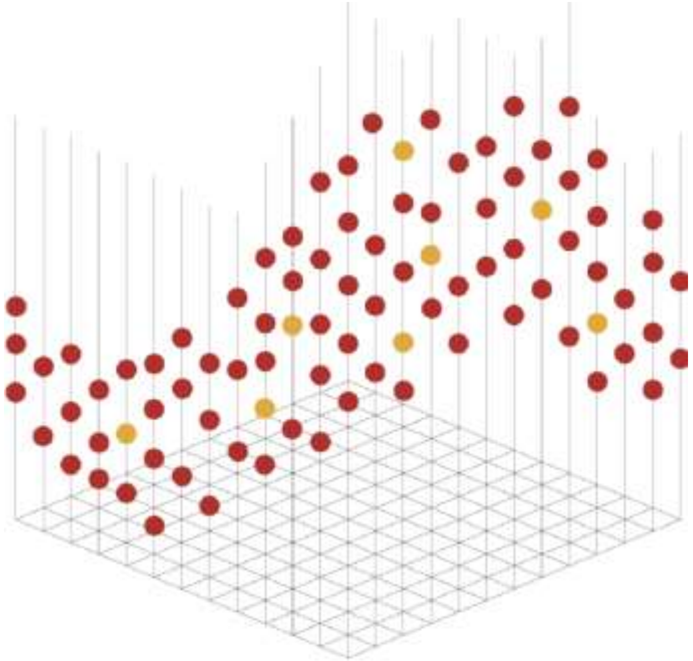
References



- [ML Extension for Sklearn: GroupTimeSeriesSplit]
(https://rasbt.github.io/mlxtend/user_guide/evaluate/GroupTimeSeriesSplit/)

- [Kaggle: Optiver - Trading at the Close]
(https://www.kaggle.com/competitions/optiver-trading-at-the-close/overview)
- [Nasdaq Closing Auction]
(https://nasdaqtrader.com/content/ETFs/closing_cross_faqs.pdf)
- [Nasdaq Stock Market Rules]
(https://www.sec.gov/files/rules/sro/nasdaq/2017/34-81188-ex5.pdf)
- [Order Book]
(https://www.investopedia.com/terms/o/order-book.asp)

BROWN
Data Science Institute



Q & A

THANK YOU!

GitHub: https://github.com/LetianY/data1030-optiver-trading-at-close/

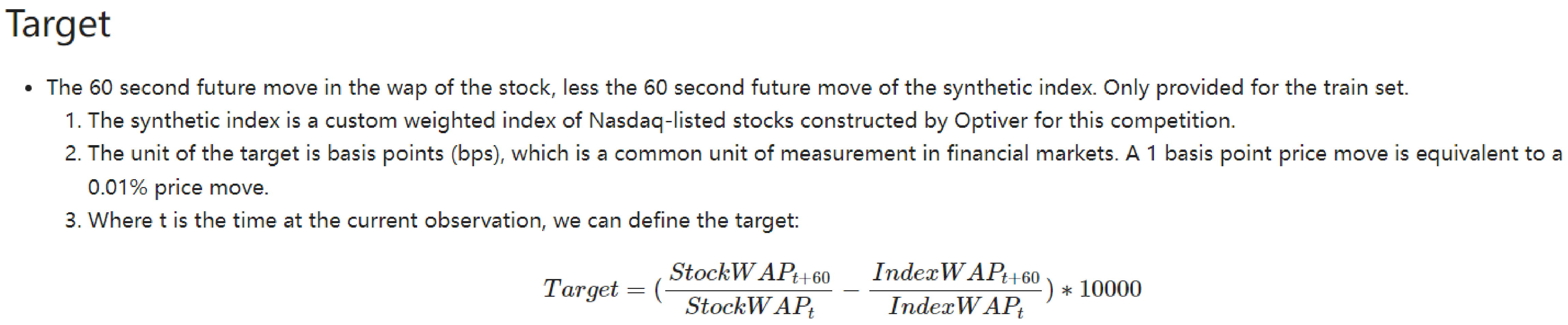
Appendix: Feature Table

Features	Description
stock_id	A unique identifier for the stock. Not all stock IDs exist in every time bucket.
date_id	A unique identifier for the date. Date IDs are sequential & consistent across all stocks.
imbalance_size	The amount unmatched at the current reference price (in USD).
imbalance_buy_sell_flag	buy-side imbalance: 1; sell-side imbalance: -1; no imbalance: 0
reference_price	The price at which paired shares are maximized, the imbalance is minimized and the distance from the bid-ask midpoint is minimized, in that order. Can also be thought of as being equal to the near price bounded between the best bid and ask price.
matched_size	The amount that can be matched at the current reference price (in USD).

Appendix: Feature Table

Features	Description
Far_price	The crossing price that will maximize the number of shares matched based on auction interest only. This calculation excludes continuous market orders.
Near_price	The crossing price that will maximize the number of shares matched based auction and continuous market orders.
Bid and ask price	Price of the most competitive buy/sell level in the non-auction book.
Bid and ask size	The dollar notional amount on the most competitive buy/sell level in the non-auction book.
wap	The weighted average price in the non-auction book.
seconds_in_bucket	The number of seconds elapsed since the beginning of the day's closing auction, always starting from 0.

Appendix: Target



Target

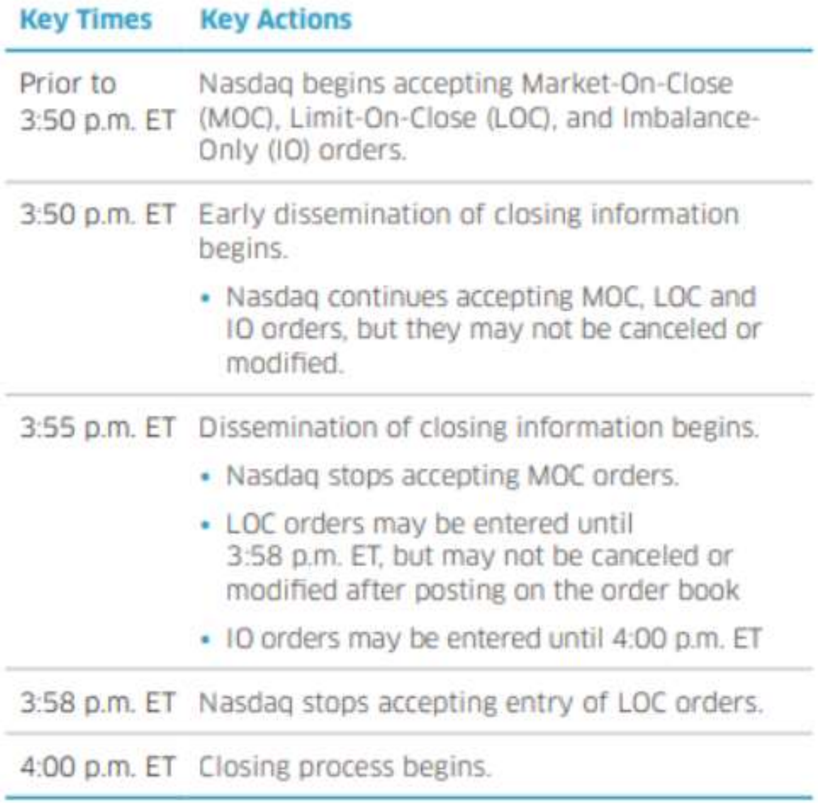
- The 60 second future move in the wap of the stock, less the 60 second future move of the synthetic index. Only provided for the train set.

1. The synthetic index is a custom weighted index of Nasdaq-listed stocks constructed by Optiver for this competition.
2. The unit of the target is basis points (bps), which is a common unit of measurement in financial markets. A 1 basis point price move is equivalent to a 0.01% price move.
3. Where t is the time at the current observation, we can define the target:

$$Target = (\frac{StockWAP_{t+60}}{StockWAP_t} - \frac{IndexWAP_{t+60}}{IndexWAP_t}) * 10000$$

Appendix: Nasdaq Stock Market

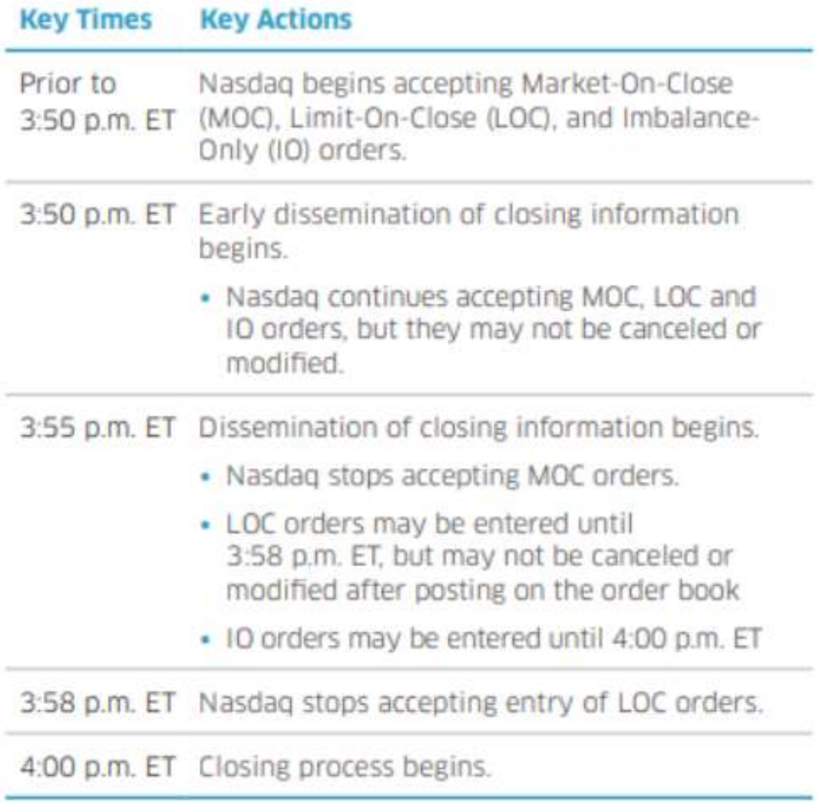
The Nasdaq Stock Market is an American stock exchange based in New York City. It is the most active stock trading venue in the US by volume. Every trading day on the Nasdaq Stock Exchange ends with a special process called the "Nasdaq Closing Cross Auction." This is a mechanism that helps determine the final or official closing price for stocks listed on the Nasdaq.



Key Times	Key Actions
Prior to 3:50 p.m. ET	Nasdaq begins accepting Market-On-Close (MOC), Limit-On-Close (LOC), and Imbalance-Only (IO) orders.
3:50 p.m. ET	Early dissemination of closing information begins. • Nasdaq continues accepting MOC, LOC and IO orders, but they may not be canceled or modified.
3:55 p.m. ET	Dissemination of closing information begins. • Nasdaq stops accepting MOC orders. • LOC orders may be entered until 3:58 p.m. ET, but may not be canceled or modified after posting on the order book • IO orders may be entered until 4:00 p.m. ET
3:58 p.m. ET	Nasdaq stops accepting entry of LOC orders.
4:00 p.m. ET	Closing process begins.

Appendix: Closing Auction

In a closing auction, orders are collected over a pre-determined timeframe and then matched at a single price determined by the buy & sell demand expressed by auction participants. For Nasdaq Closing auctions, the exchange begins accepting orders at the start of the trading day and begins publishing the state of the auction book at 3:50pm ET for 10 minutes before the market closes at 4pm ET, at which point the orders are matched instantly at a single price.

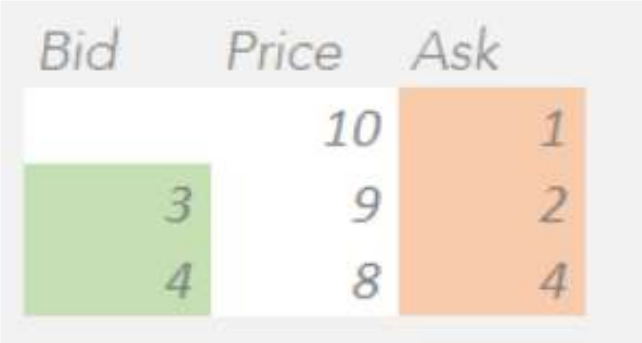


Key Times	Key Actions
Prior to 3:50 p.m. ET	Nasdaq begins accepting Market-On-Close (MOC), Limit-On-Close (LOC), and Imbalance-Only (IO) orders.
3:50 p.m. ET	Early dissemination of closing information begins. • Nasdaq continues accepting MOC, LOC and IO orders, but they may not be canceled or modified.
3:55 p.m. ET	Dissemination of closing information begins. • Nasdaq stops accepting MOC orders. • LOC orders may be entered until 3:58 p.m. ET, but may not be canceled or modified after posting on the order book • IO orders may be entered until 4:00 p.m. ET
3:58 p.m. ET	Nasdaq stops accepting entry of LOC orders.
4:00 p.m. ET	Closing process begins.

Appendix: Book

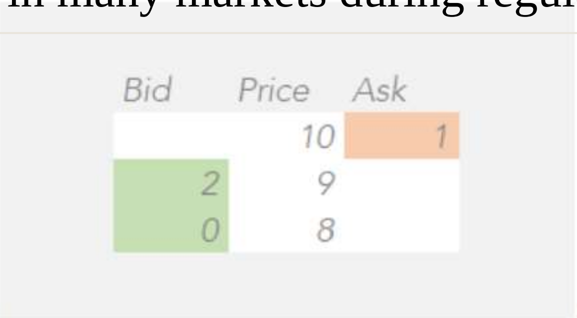
Auction Book: This contains orders that are executed through an auction mechanism. In auctions, buy and sell orders are aggregated, and a single price (the auction price) is determined where the maximum volume can be executed. Auctions are typically used at the opening and closing of markets, though some markets may have intraday auctions as well.

Non-Auction Book: This typically contains orders that are executed continuously during trading hours outside of the auction mechanisms. They are matched on a continuous basis as and when compatible buy and sell orders (in terms of price and other conditions) are entered. This is the usual method of trading in many markets during regular hours.



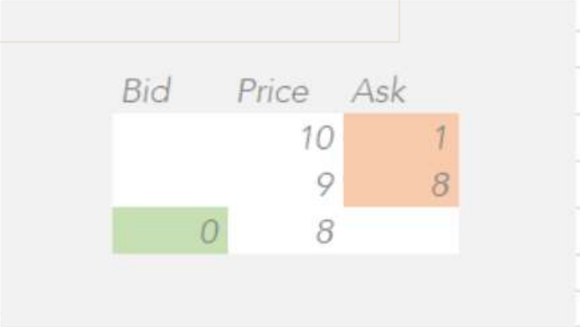
Bid	Price	Ask
	10	1
3	9	2
4	8	4

Before ask:



Bid	Price	Ask
	10	1
2	9	
0	8	

After an ask of 10 shares of price 9:

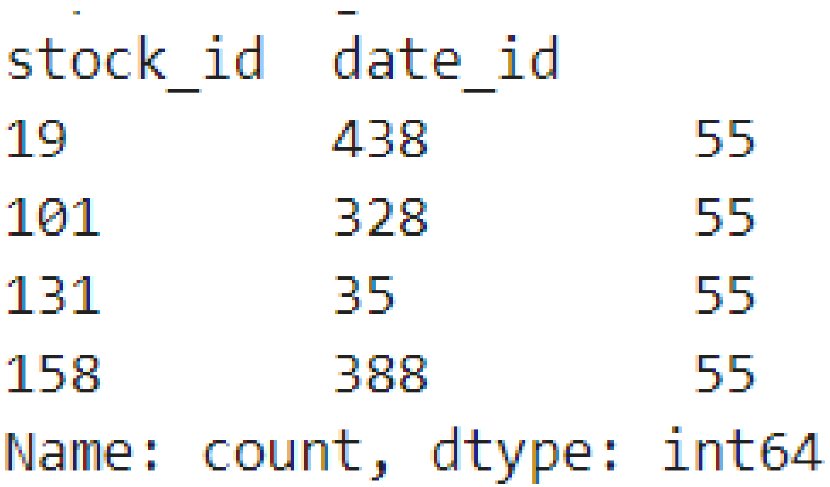


Bid	Price	Ask
	10	1
	9	8
0	8	

Appendix: Missing Values in Time Span

	Missing Date Count	Missing Date Min	Missing Date Max
stock_id			
69	37	0	36
73	1	320	320
78	4	0	3
79	181	0	180
99	1	138	138
102	295	0	294
135	191	0	190
150	59	0	58
153	70	0	69
156	37	0	36
199	88	0	87

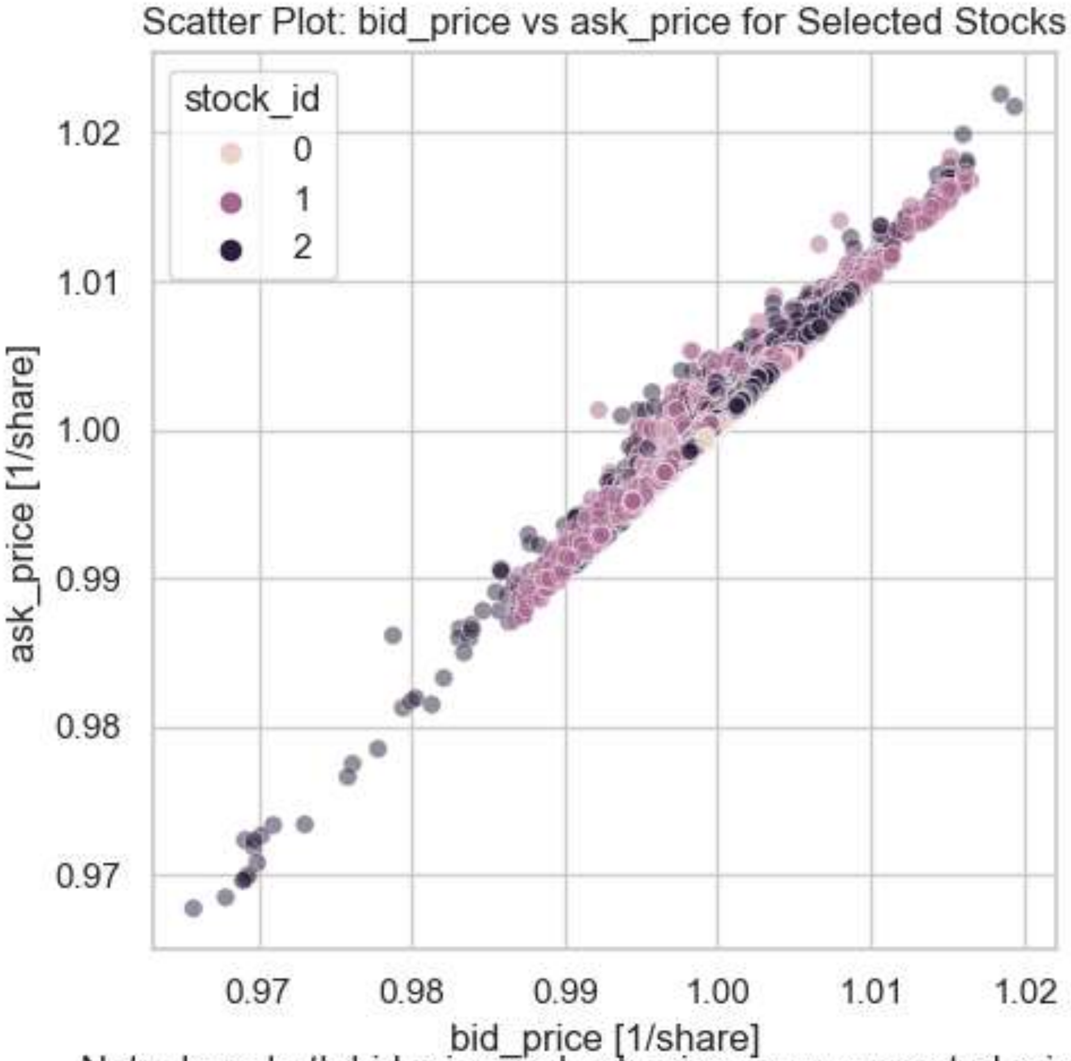
Appendix: Which stocks are missing 220 values in wap



```
stock_id  date_id
19        438        55
101       328        55
131       35         55
158       388        55
Name: count, dtype: int64
```

Appendix: Bid-Ask

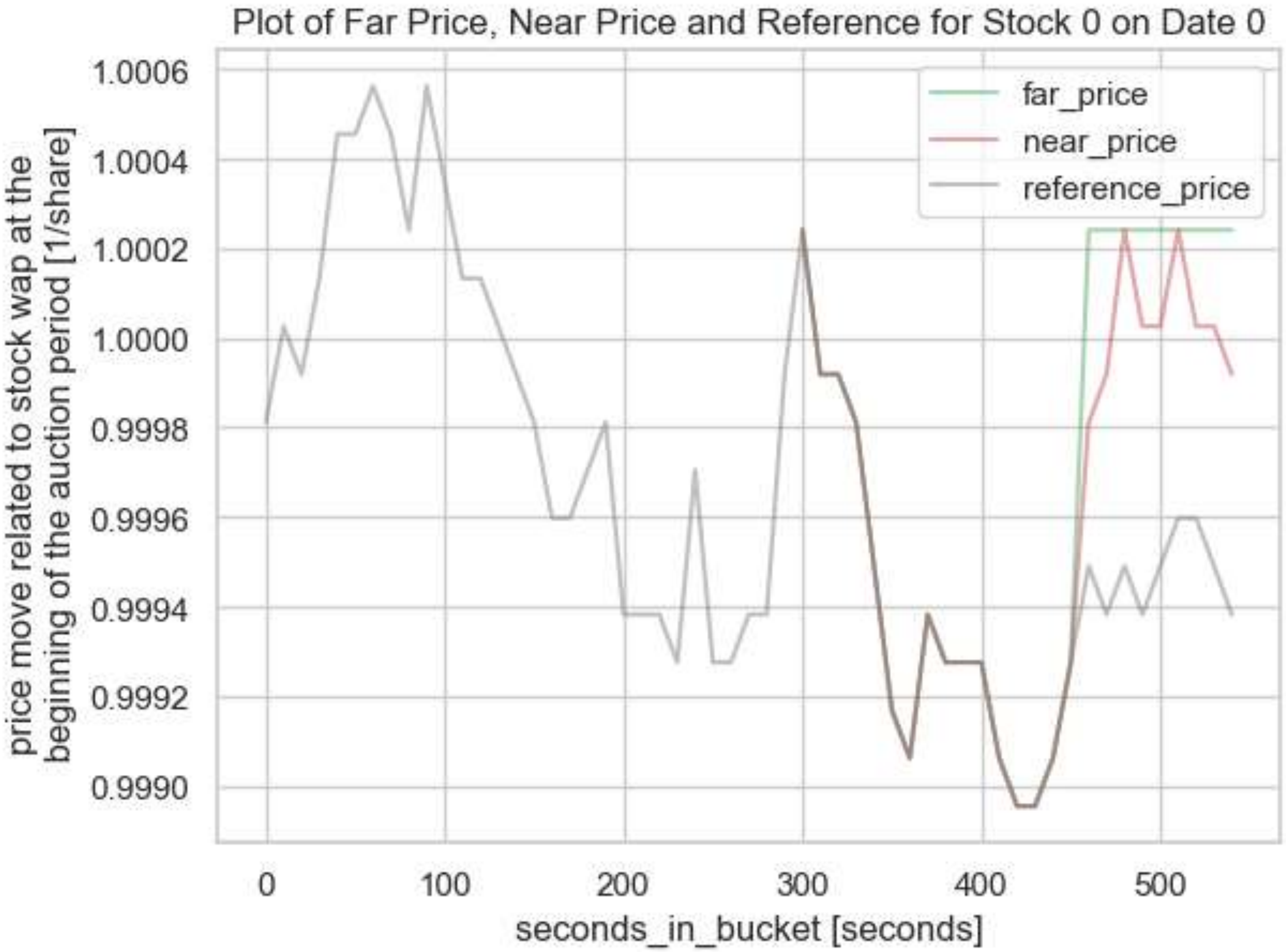
Bid-ask price scatter plot for individual stocks



Note: here both bid price and ask price are a converted price move related to stock wap at the beginning of the auction period.

Appendix: Far-Near

Far price and near price only starts in the last 5 mins of auction


Plot of Far Price, Near Price and Reference for Stock 0 on Date 0
price move related to stock wap at the beginning of the auction period [1/share]
seconds_in_bucket [seconds]
far_price
near_price
reference_price