- We conducted exploratory analysis to familiarize ourselves with the data and decide which direction to move in.

- Commented the starter notebook to understand what they did.

- Determined number of stocks, individual stock\_ids (57 and others are missing. This matters. 122 total but highest id is 126), # of time ids, time id values, 600 seconds per bucket (10 minutes) and all time ids are consistent throughout all stocks

**- Defined functions to easily extract data at every level (book/trade, stock id, time id, seconds in bucket).**

**- Defined functions to visualize different variable interactions for any given stock id time id pairing**

**- Amt of trades executed throughout bucket**

**- Bid size, ask size throughout bucket**

**- Bid price, ask price throughout bucket**

**- Big-ask spread volume with executed trades and trade sizes throughout bucket**

Next steps:

- Aggregate Features: For each stock and time\_id pairing, extract aggregated features such as average price, total volume, volatility, spread between bid and ask prices, etc.

**- Reverse Engineering TimeID**

- Feature Engineering: Derive new features from the existing data columns. Some potential features to consider:

- Moving averages of prices.

- Momentum and rate of change indicators.

- Order book imbalance (difference between bid and ask volumes).

- Rolling volatility or standard deviation.

- Price jumps or gaps between consecutive time buckets.

- K-means Clustering: Use the K-means algorithm to cluster time\_ids based on the extracted features to identify similar market conditions.

- Reverse Engineer time\_ids: Investigate the exact dates and times for each stock and time\_id pairing. Integrate with historical data from platforms like Yahoo Finance.

Model Selection:

* Start with K-NN
* Include for Milestone 3 prof. Hamm’s thoughts on where we should go with Models