**MaDCaAT**

Market Data Collection and Analysis Tool

**The case against current market data charts**

Typical market data analysis tools have lost their shine with the heightened understanding of statistics provided by the MIDS program. Specifically, the simple price chart that is the go-to for purchase and sale activity by the average investor. What does this chart of Apple (AAPL) really tell us?



It looks like the prices have been rising, but note that the price charted is just the *closing* price. It is simply the price of the last transaction of the day. The market votes with its dollars, and if you were to accept this chart as a fair valuation of Apple over time, you might also accept that a restaurant’s quality can be determined by only the last daily review over a period of time on Yelp. Or the last Tweets per day on a particular subject are enough of a sample to judge the sentiment. Or the quality of a class should be based solely on the last reviews thoughts per day.

Consider these two facts about typical price charts:

1. The prices that make up the chart are not even randomly sampled from the day.
2. Transactions are typically for between 100 and 2000 shares. A day’s volume of APPL may be over 10 million shares, meaning the last transaction probably doesn’t even represent 1/100th of a percent of the total day’s activity.

How can one really identify a trend from such incomplete data? From a series of closing sales? This is a little data view of a big data problem.

**The case for a statistical view:**

Each trading day provides an abundance of information that can’t be usefully summarized by closing price and a few statistics. Instead, we can study all the trading activity occurring per day by calculating a list of prices vs. sizes of transaction. At the end of the trading day, we will have a histogram showing the number of shares transacted at each price point throughout the day. We can find basic statistics about that trading day such as the mean price per share and the standard deviation. We can study the distribution of prices to see how much they differ from a normal distribution, and how much they differ from other trading days.

Further, we can collect this information over time and produce meta statistics about trading days. What is the average standard deviation over the past few days? How does the variance of the past ten trading days look? Can we create a liner regression using past daily distributions to suggest a future distribution? What does a regression of mean prices look like? Are there values that remain relatively constant over periods of time that can suggest buying or selling opportunities?

Each of these statistical features provides a toe-hold to make a decision. They allow us to analyze the price behavior within a clearly understood statistical framework. There may still be unintelligible signals and results not complete enough to make a decision, but there is more information in this approach than in a price simple chart.

**MaDCaAt**

Madcat is half library and half psudo-real-time analytics tool. MadCat stores historic daily sales activity for a number of symbols and can calculate:

* Daily Histogram.
* Essential descriptive statistics of both price and size information.
* Differences in daily statistics (z test)
* Probability distributions based on historic prices
* Meta statistics
  + Descriptive statistics over time
  + Z test over time
* Regressions (of essential descriptive statistics)

Using python interface, a user can formulate statistical theories about market behavior. These theories can be tested by observing a histogram constructed in psudo-real time from a live market data feed. For example, we can construct a histogram of AAPL price and size during the trading day, as trades occur. The experimenter may theorize a probability distribution of today’s mean price and total volume and make trades according to the market’s response (as visualized by the histogram) to those parameters.