Byte Quant Researcher Challenge

**Created**: 2021-01-15

**Author**: Jan Hoekman

#### Context

For this challenge we are exploring tweets of selected investor accounts covering the period of 2019/2020. We’d like to analyze the raw tweets, make some observations, form hypotheses, and outline a plan to create a strategy we could backtest as a next step.

#### Objective

Overall I’d like to get a grasp of how *you approach your research*: your code, your ideas & general methodology. As we’re limited in time, data, and tooling due to this being a contrived hiring challenge you should especially also explain how you’d do it differently if we wouldn’t have these constraints / in the future when working together. I’d encourage you to spend at 1-2 hours on this only.

As such, the idea is to:

1. Come up with a few insights

2. What would you do with more time? / What different tooling and/or data would you like / develop yourself?

3. How would you turn this into “backtestable” strategy & how would it look like (e.g. what type of strategy is it?)

#### Data

The dataset “tweets.csv” contains the tweets. Tweets can contain multiple symbols, but for keeping it simple here we only analyze the data of the first mentioned symbol: column “first\_symbol”. I’ve also only added a small subset of all the tweets we track to keep it manageable & quick.

In addition to various features that come with each tweet (such as whether it’s a retweet, hashtags mentioned etc.) I’ve also added two columns that proxy the market impact of a tweet by measuring the volume and price in the post-event 5 minute interval (including the minute the tweet occurred in) and comparing it to the preceding 30 minute interval using a normal gaussian distribution. The columns are “sigma\_price” and “sigma\_volume”.

To enable the exploration of these tweets in conjunction with further market data I’ve included market data for all “first” symbols contained in the dataset. The data is stored as parquet files.

As a starting point I have created a jupyter notebook for you in the drive, which for instance shows you how to load the market data.

#### Deliverables

* Notebook at least a few hours before our meeting
* In our meeting we can discuss everything and talk about this challenge, Python & other tooling, and how we would design a better research process together.

If you have any questions contact me via email.