Data Ingestion Report

Lubin Sun | ls6211@nyu.edu

This is the trading data of Bitcoin price and volatility in minute frequency. Bitcoin market are driven by information asymmetries and the injection of new information in trades into market prices. This dataset has a limit order book file and market price file. I processed the market price file (“bitcoin.csv”).

Data Ingestion:
Index: market timestamp, it has the form like “2011-12-31T09:20:00Z” of unclear type. I converted it to the standard timestamp type.
Open: the open price of the 1 minute interval
High: the highest price of the 1 minute interval
Low: the lowest price of the 1 minute interval
Close: the close price of the 1 minute interval
Volume_(BTC): how many Bitcoin is traded in 1 minute interval
Volume_(Currency): how many US dollar is traded in 1 minute interval
Weighted_Price: Volume Weighted Average Price (VWAP), the weighted average price using the volume as weights

Statistics:

summary	Open	High	Low	Close	btc_volume	usd_volume	Weighted_Price	return	Volatility
count	4363456	4363456	4363456	4363456	4363456	4363456	4363456	4363456	4363456
mean	2751.7300198774400	2753.7034329212800	2749.6006396288640	2751.6866227131400	9.952766746924490	21047.603940597300	2751.668008361300	9.24468562569029E-05	4.1027936353659100
stddev	3686.989393221720	3690.154815717840	3683.503242669680	3686.908572668040	31.02048398406251	86491.64481331730	3686.882403709020	0.18593490320432400	10.417044131357500
min	3.8	3.8	1.5	1.5	0.0	0.0	3.8	-0.9974514068234340	0.0
max	19665.76	19666.0	19649.96	19665.75	5853.852165913	7569437.0613	19663.298888	388.3733333333330	729.0100000000000

- Return has a mean near 0, which makes sense because the return is a random walk.
- Volatility has 0 min, which means in that minute, the price is not changed. This can happen in 2011, back then Bitcoin was thinly traded.
- Open/High/Low/Close data are consistent.

Conclusion:
In this project, we are looking for how the news in Twitter and Reddit can affect the price of Bitcoin. We assume that effective news on Twitter and Reddit can lead to significant price change. Thus we need two new features: return and volatility.

Return: close - lag(close, 1) / lag(close, 1)
Volatility: high - low
Since price can be driven by multiple factors. The trading activity like short squeezing can also lead to price sharp changes. We use these two features to justify if a price change is caused by news. If the volatility is high in one interval but the return does not change significantly, it might be a trading-caused event. If the return is changed significantly and the volatility clusters in a longer period of time, this is probability a news-driven event.
With these two features, we can make a tentative research on how the events impact the Bitcoin market.

Potential Improvement:
The asymmetry in limit order book (LBO) can also be a pattern of different types of events. I will add that file into the project at a later stage. Also we can add more pattern to justify or validate our research