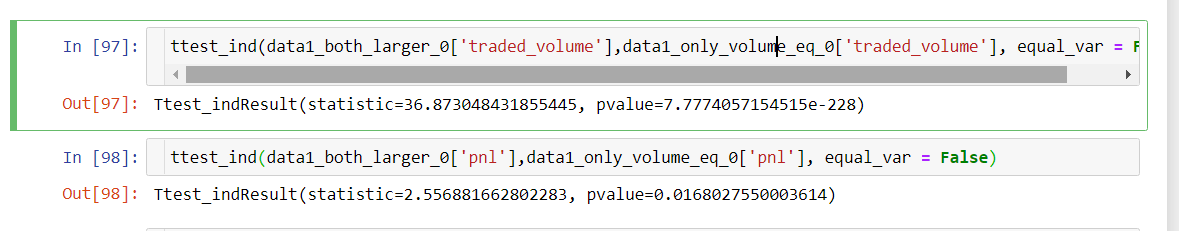
Before reading this report, I would like to stress that I have not written production-oriented script for the test. It is mainly due to the fact that I am not 100% confident that my analysis is on the right track.

We will investigate 2 types of anomalies in this writeup.

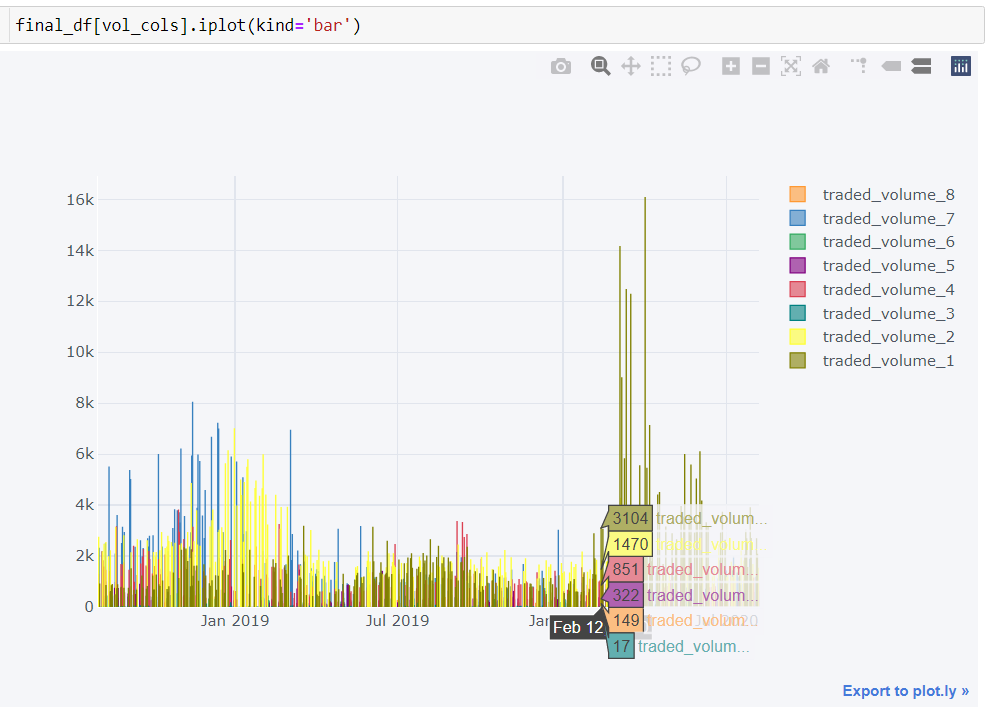
1. Data feed missing (could be from upstream system)
2. Data being outlier
3. Data feed missing
4. For exchange 12, for 20 business days we are seeing no data feed from 2019-07-04 to 2019-07-31.
5. For row 2725 on 2018-12-12 and row 4433 on 2019-02-11, traded\_volume is not 0 but pnl is 0.0. Might want to flag pnl in case upstream breaks or not.
6. Data being outlier:
7. Group our data into 2 groups. *Group A* is those both volume and pnl >0. *Group B* is those volume is 0 but pnl is not 0. They might represent 2 types of trades. Perform a t-test here to see whether this hypothesis holds. 

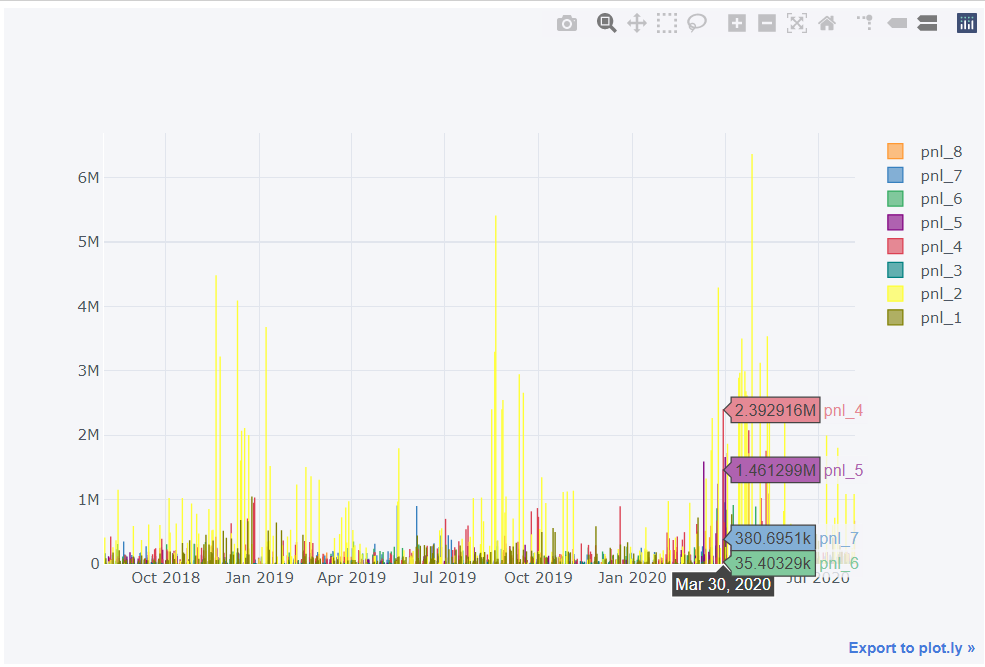
Both shows statistical significance. Will treat two groups differently.

1. For our test, let’s continue with *Group A*.

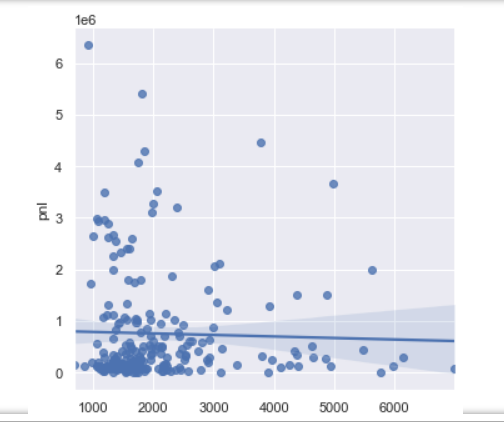
Since we do not know how to decide to route to these 13 exchanges, we need to see whether their pnl/traded\_volume is comparable.

Looks like for some exchanges it’s quite true.



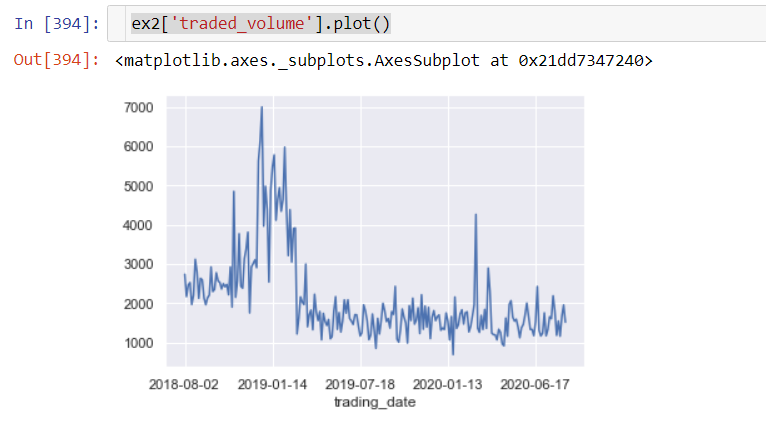


1. We will try to focus on 1 single time series and do some analysis. Use exchange 2 for example.

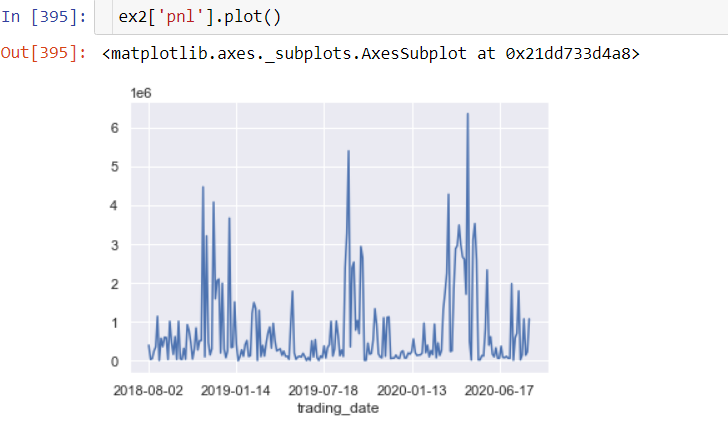


No linear correlation spotted for pnl and traded volume.

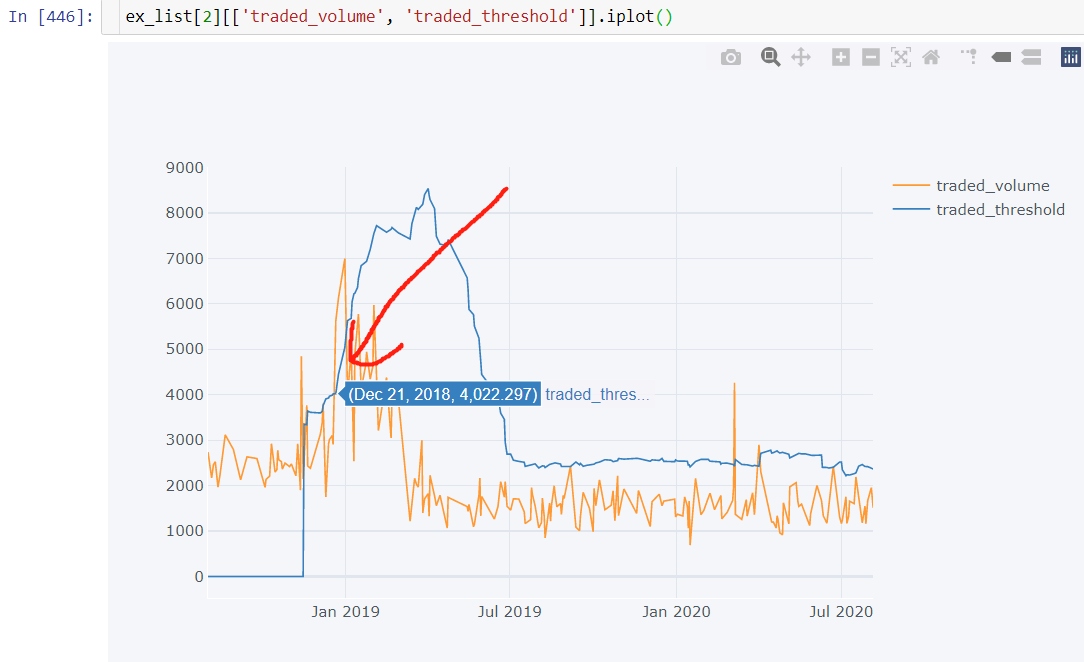
Clearly, volume has paradigm shift:



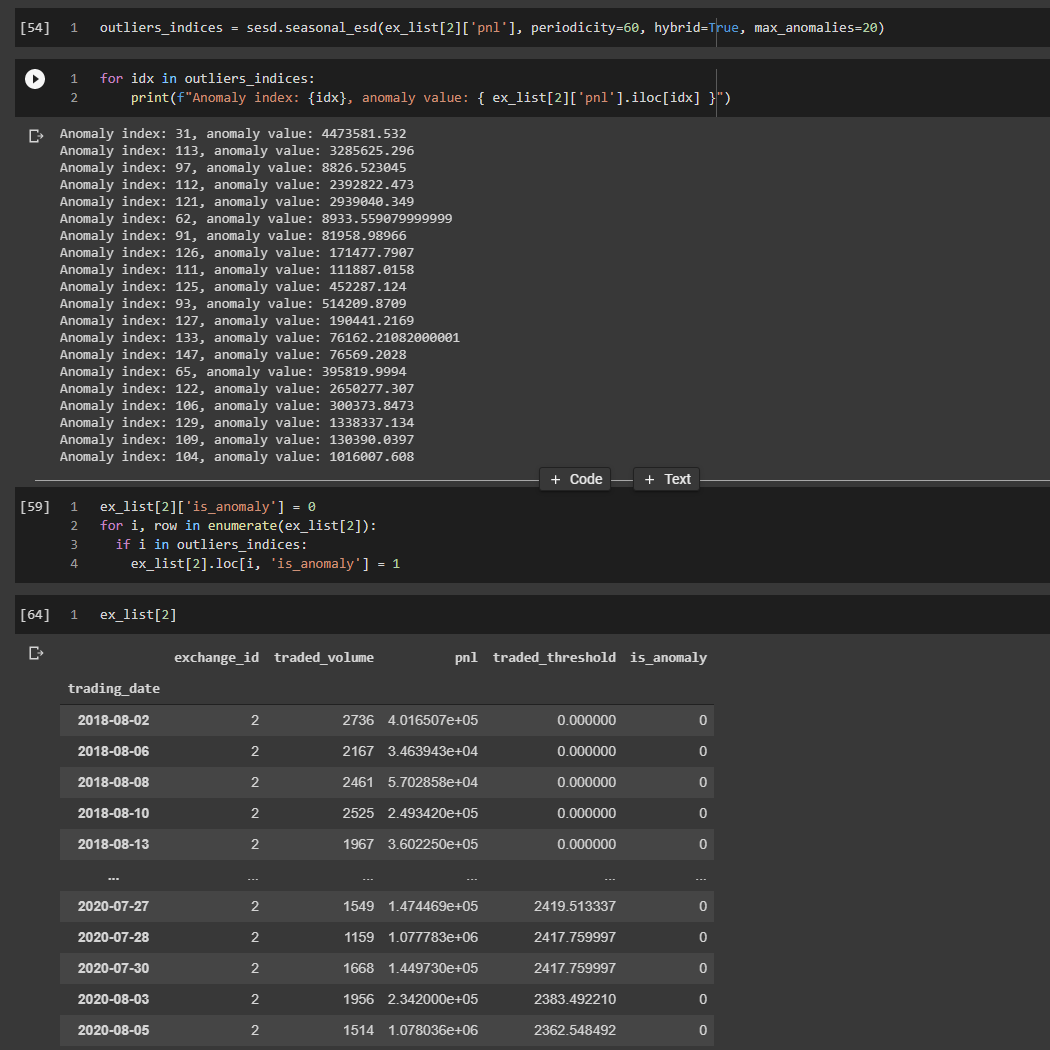
Pnl look more periodic/seasonal. Also interestingly all positive.



For traded volume, we do a rolling window with median and standard (real-time threshold) deviation. Could use MAD/IQR as well. Intervene/getting label happens at red arrow.



For PnL, we do a retrospective version.



References:

