Justin Holladay, 2021
Cryptocurrency Exchange Pump and Dump Detection

Context:
The quick rise of interest in cryptocurrency has led to high volumes of money traded on unregulated markets. Many kinds of manipulation exist in these markets; a Pump and Dump is one. The concept is that an insider group buys an asset, then convinces many outsiders to buy the asset to drive up the price quickly. At that point, the original insider group will sell their assets, effectively draining the outsider group's investment.

Many public and private pump and dump groups exist on messaging platforms like Telegram and Discord. These groups have member counts upward of 150,000 and coordinate pumps frequently. This project seeks to identify this type of market manipulation using price, volume, and transaction data.

Problem
How can we identify Pump and Dump market manipulation on centralized cryptocurrency exchanges?

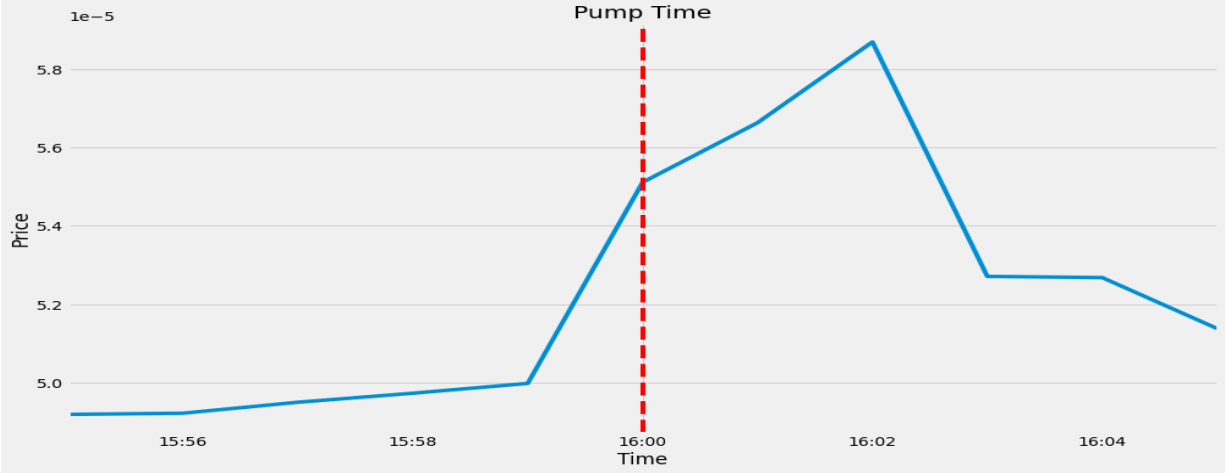
Wrangling
Trading data was gathered from many different assets from the Binance exchange. Previous work by SystemsLab [0] provided time, data, and asset name until Q1 of 2021 of pumps advertised on telegram groups. I gathered additional data for this capstone project through Telegram and Discord.

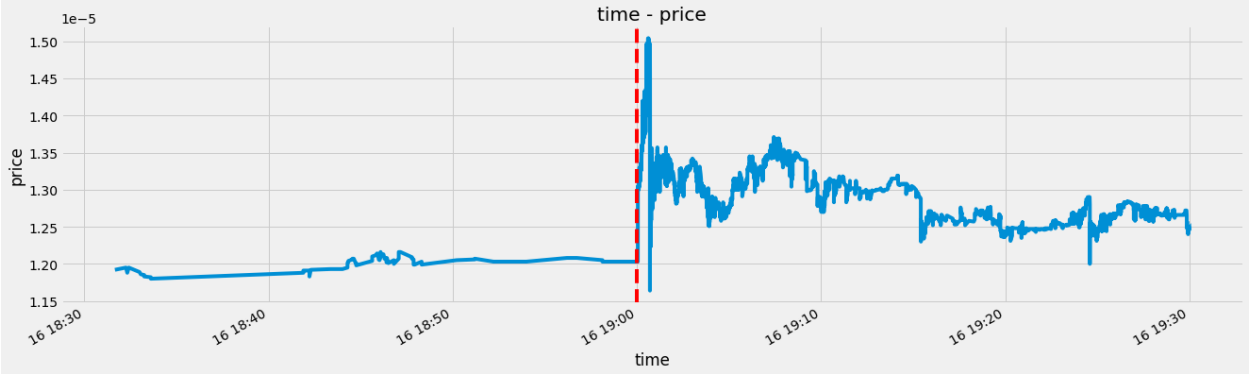
A script was created from the gathered asset names and pump times to download K-line and trading data of 400 instances from the official Binance Github. Raw data of the whole month of the pump was downloaded and K-lines at 1m resolution. The raw data was then labeled and time indexed for use with pandas.

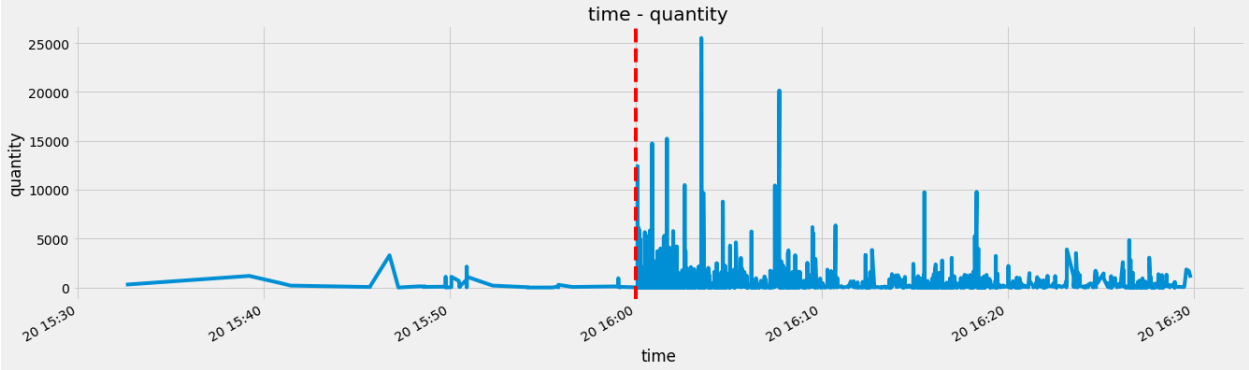
Exploration
Initially, exploring K-line data for the assets showed a slight variation in price and volume for the month. When looking at the target day and time, there is an expected sudden spike in volume, number of quotes, and price when the asset is announced. In many cases, there is a smaller spike several minutes or seconds before the announcement of the asset.

- All coins are under a 50,000,000 market cap.

- Successful events will increase the price 2x-5x on average.
- The number of individual quotes increases 10x, indicating more participants in the scheme.







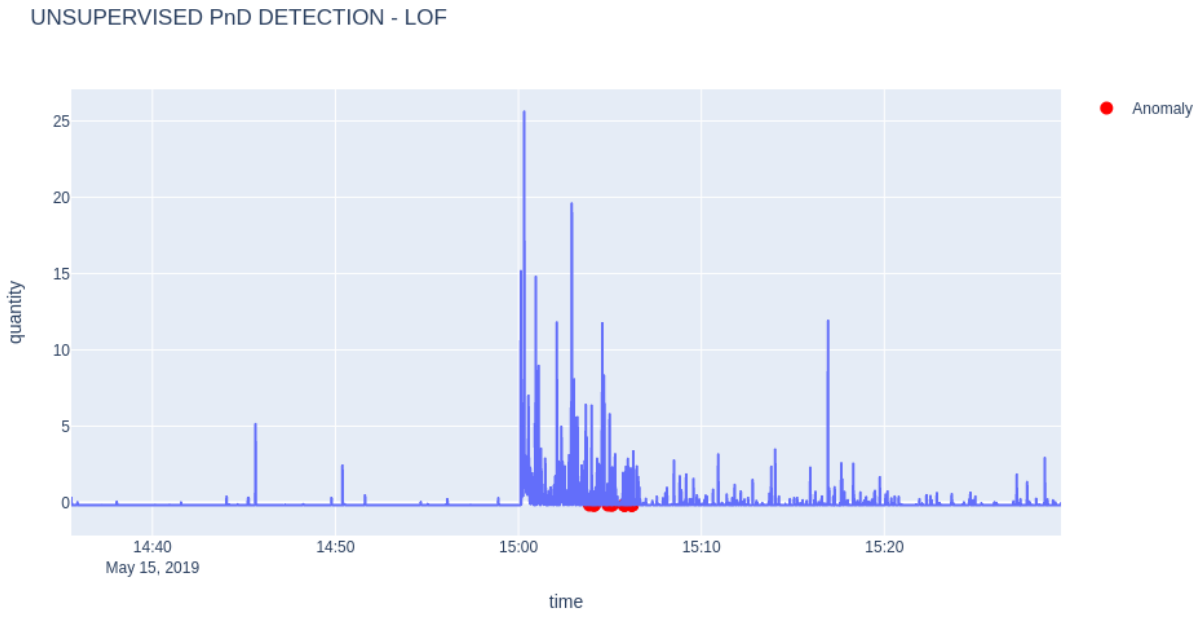
Preprocessing

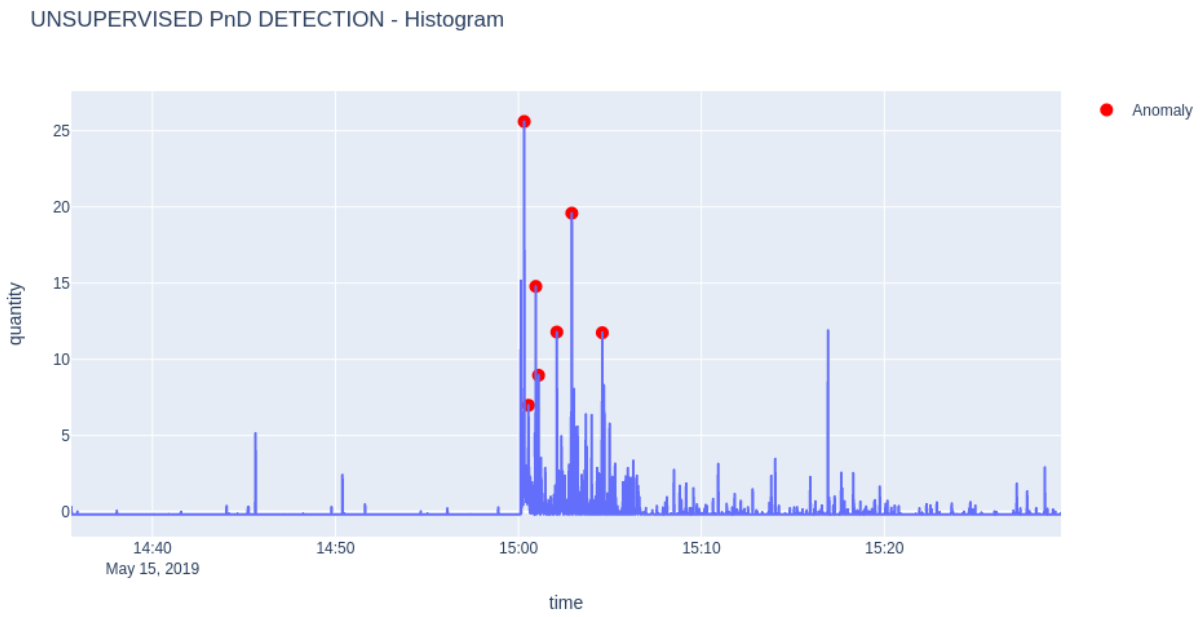
This project aims to detect manipulation in a higher resolution than 1m for the window of time we are exploring. K-line data was discarded, and trading data was used. This data contains the specifics of each trading order and when it was placed, up to the millisecond.

After exploration, data from 15 minutes before and after the scheduled pump was extracted from each instance.
The trading data was gathered from each CSV file and resampled into 1s bins. Each column was appropriately aggregated and standardized. Any missing columns were forward-filled, and no columns were dropped. All were merged into a single dataset.

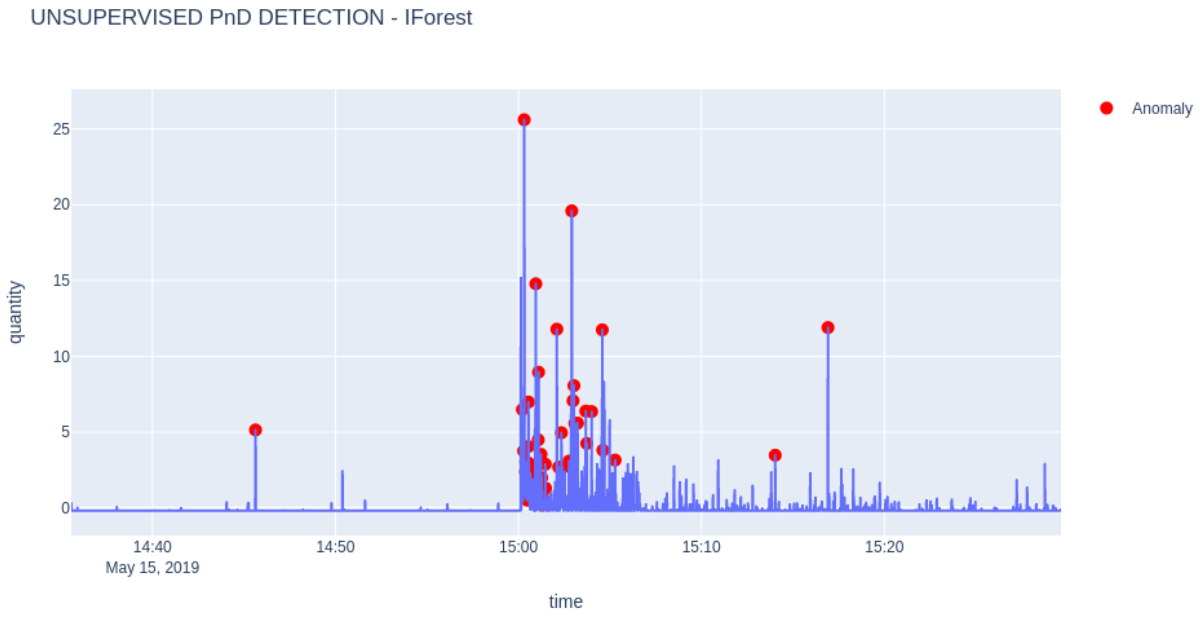
Modeling:
Principal Component Analysis, Local Outler Factor, and Isolation Forest were chosen for modeling Testing.

LOF and Histogram models could detect very few anomalies before the scheduled pump time, though and defined inconsistent anomalies at the pump time.





Isolation forest was chosen because our target anomalies may not have a standard value distribution. Inspection of IForest results shows it aligns better with the cycle of the scheme. It is discovering some orders with much less volume that can be assumed as the insiders.



Conclusion:

Applications of this exploration can be used for exchanges, businesses, and investors directly. Including:

- Exchanges: Identify participants fo the scheme
- Business: Protecting clients from identifying higher-risk manipulated assets. As well as setting bot activity.
- Individuals: Keeping them cautious of swing trading risky assets and investing in the schemes.

Extensions on this research could include exploring more features like inferring the number of market vs limit orders set. As well as adjusting the window of time to use and the frequency of resampling.