Stonk Stats

Hello, Superstonk

I’ve been working on something that may be of interest, especially to apes with a wrinkle or two in statistics and data. I would appreciate comments, criticisms, and suggestions as to how to improve what I’ve done. The code I’ve written to reproduce everything here is uploaded to https://github.com/kirpi-1/stocks, so feel free to clone/fork and check the work.

Problem

Our beloved stonk seems to behave very strangely, at times seemingly 100% correlated with totally unrelated stocks. Various theories have been proposed, including swaps and baskets and other things I don’t understand because my brain is as smooth as a b-baby’s bottom when it comes to financial things. I do know a little coding and statistics, however, and was curious if I could find anything suspicious.

Premises

1. The stock market has some base correlation, i.e. in good times all the stonks go up, in bad times all the stonks go down.
2. Two randomly picked stocks that have nothing in common should only be correlated insomuch that the general stock market is correlated
3. Two randomly picked stocks within the same sector should probably be more correlated, but only insomuch as stocks in that area are correlated
4. Anything that shows unexpected correlations could have something in common, a common driving force (e.g. reddit/social media, hedgefund fuckery, etc.)
5. Correlations probably change over time as trading strategies change

So, the basic idea is to examine correlations over time between any two stonks (usually GME and another stonk) and look for periods of unusually strong correlation. If two stonks spend more time than is expected at strong correlation, then there might be something unusual about the pair.

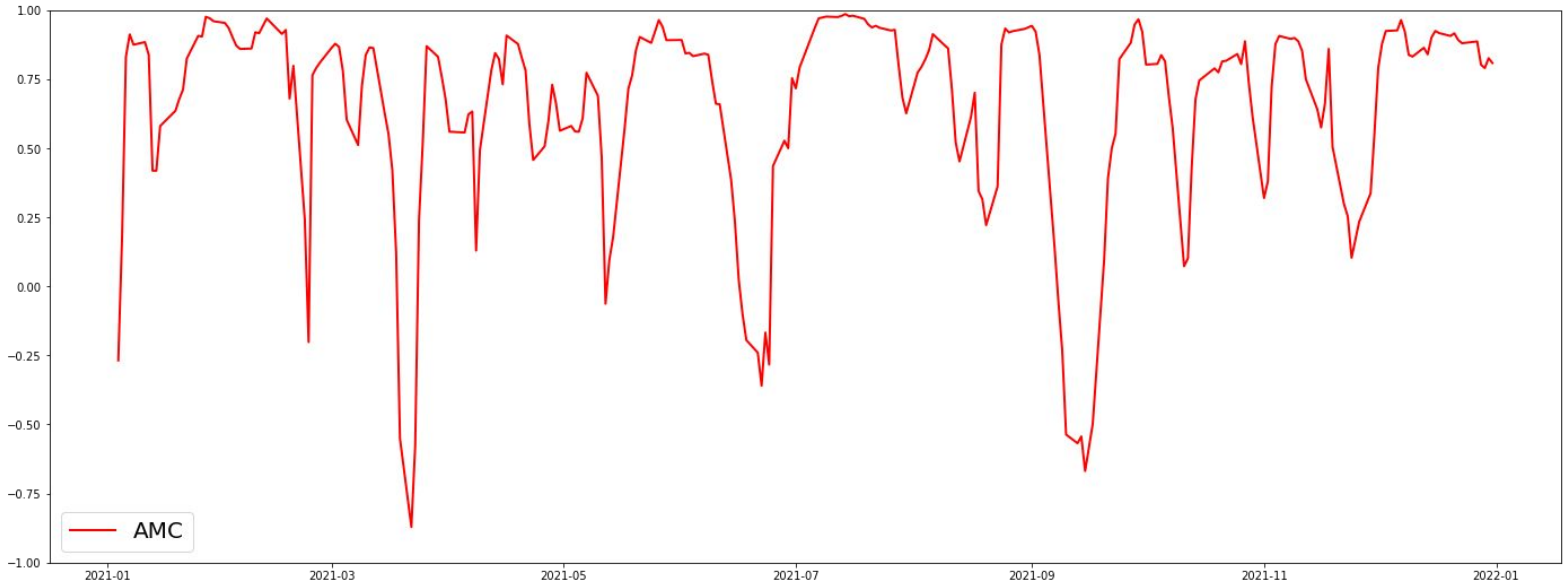
Methods

1. For a set of stonks (I combined the S&P 500, Russell 1000, and some random stuff I thought might be interesting), calculate the moving correlation time-series between all of them for a set period of time.
2. For every pair of stonks, calculate the distribution of correlations. Calculate the global distribution and within-sector distributions as well for comparison.
3. Pick a target stock (GME). Look at the distributions of correlation with every other stock. Any stock that differs too much from the global norm is suspicious. We may try to measure the difference in distributions by something like the Kolmogorov-Smirnoff test, although there are other methods.

Hypotheses

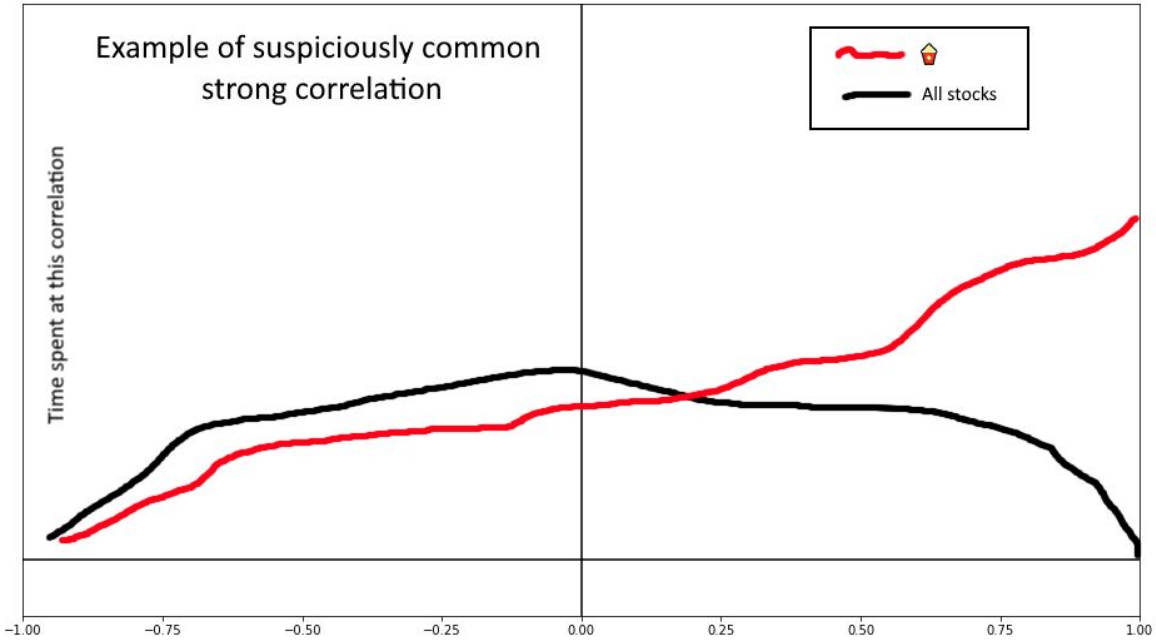
1. In general, the market as a whole should have zero or slightly positive correlation. If every stock were independent, should be a normal distribution around 0. However, I'm guessing stocks are highly dependent on each other, so maybe a flat or something slightly positive.
2. Stocks that have something in common with GameStop should be strongly correlated one way or the other
3. Things like popcorn and towel should spend a lot more time than expected with strong correlation values
4. Meme stocks are probably correlated among themselves due to retail trading them together, i.e. if apes and regards get wind of some stocks on reddit, they may be buying/selling all of them together.

10-day Rolling Correlation of 🍿's closing price



Popcorn spent an awful lot of time during 2021 being extremely strongly correlated with GME. This is what we're looking for.

Example of suspiciously common strong correlation



X-axis is correlation value. Y-axis is a count of how many time periods were spent at a particular correlation value. Thick black line is calculated for all stocks as a reference.

If a stock doesn't look the average stock, it is suspicious.

Because I'm interested in strong correlations (>0.5 or <-.5), I want to also specifically find those that spend a lot of time at the top and bottom. More on that later.

This is how I want to find other suspicious stocks.

Results

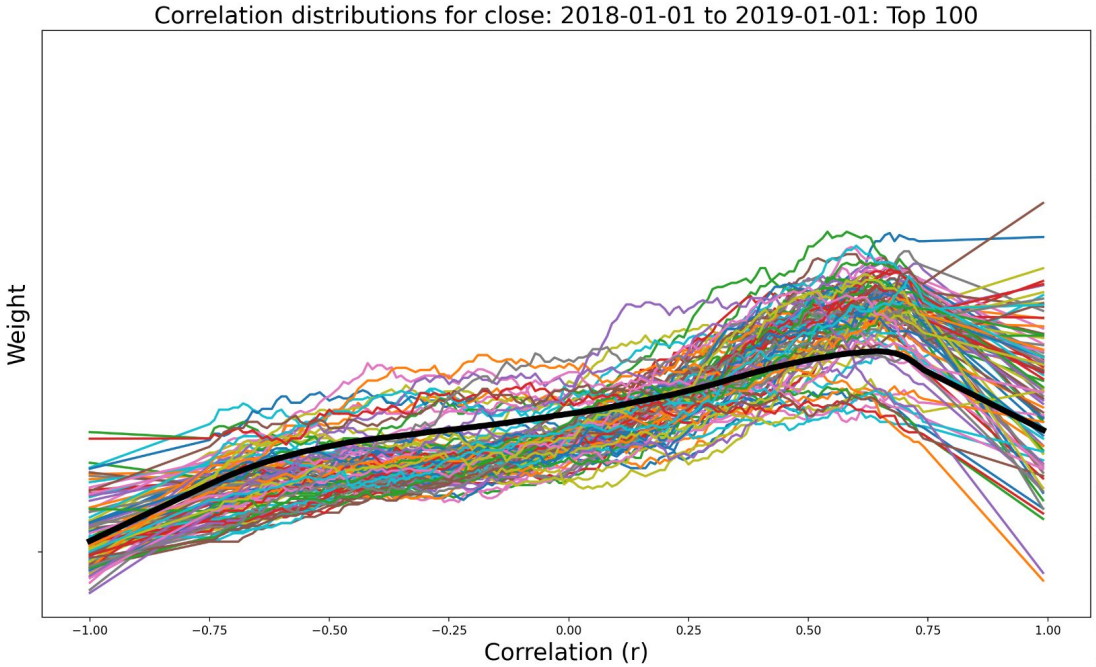
My first instinct was to use closing price, although other values are possible (gain, volume, high-low, etc.).

When looking for outliers, I am using two main criteria:

1. Kolmogorov-Smirnoff test. This gives a sense of how close two distributions are to each other. I'm using it to compare a single stock's correlation distribution to the global correlation distribution
2. Something I'm calling "Rolling Score", which I'm calculating by summing all the correlation values when they're over/under a certain range. In this case, I'm counting up all the correlations that are more extreme than +-0.5, which will give me a positive score and negative score. These should find the stocks that spend the most time strongly correlated with GME

I first began with 10-day rolling correlation over a 1 year period, starting on January

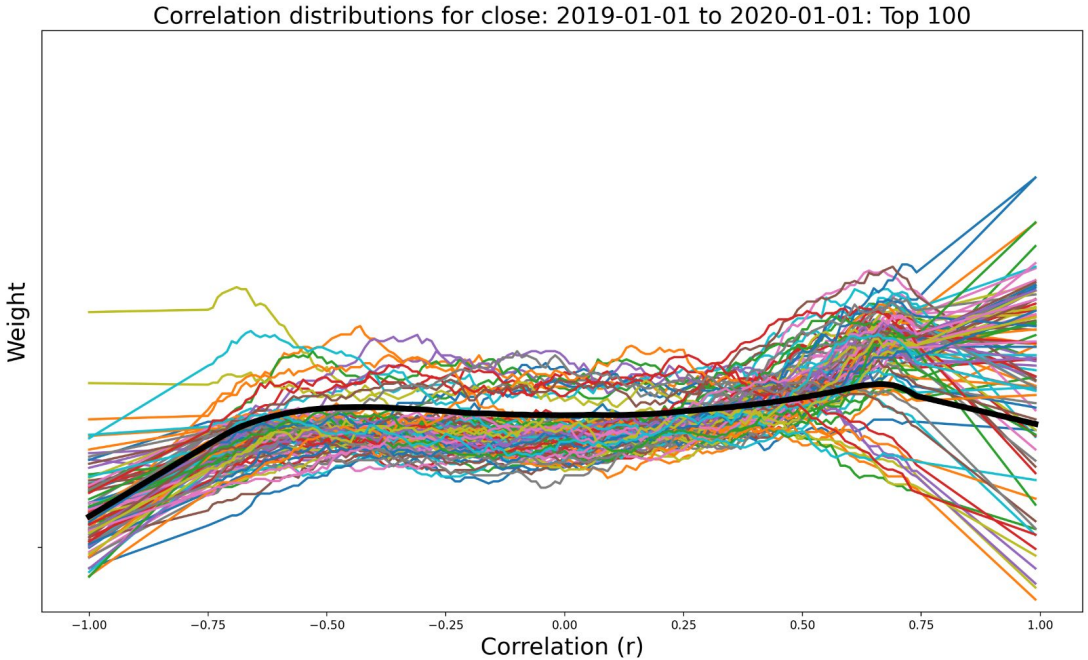
Results - 2018



There are a few standout stocks, and would you look at that, our friends towel and great purchase are there. This predates anything on reddit with GME because I don't think DFV found GME until 2019.

pos_score		neg_score	
BBY	112.90	ENPH	-53.46
JWN	102.60	EXEL	-51.42
MANH	100.20	AGR	-47.64
TMUS	96.10	GE	-46.71
EFX	94.08	MKTX	-45.78
HBI	92.45	BX	-44.52
BC	91.86	AEE	-44.36
SRPT	91.67	NEM	-44.33
GPS	91.23	DZSI	-43.85
BBBY	91.09	FWONK	-43.76
CMG	90.09	PEG	-43.53
FBHS	89.83	EVRG	-43.16
FNB	89.21	FWONA	-42.49
POST	89.17	WHF	-42.27
VAC	89.02	RGLD	-42.24
CRI	88.87	EXC	-42.16
CASY	88.62	TDG	-41.34
WBA	88.42	OGE	-41.10
LH	88.41	DTE	-40.98
LEG	88.36	RNR	-40.73

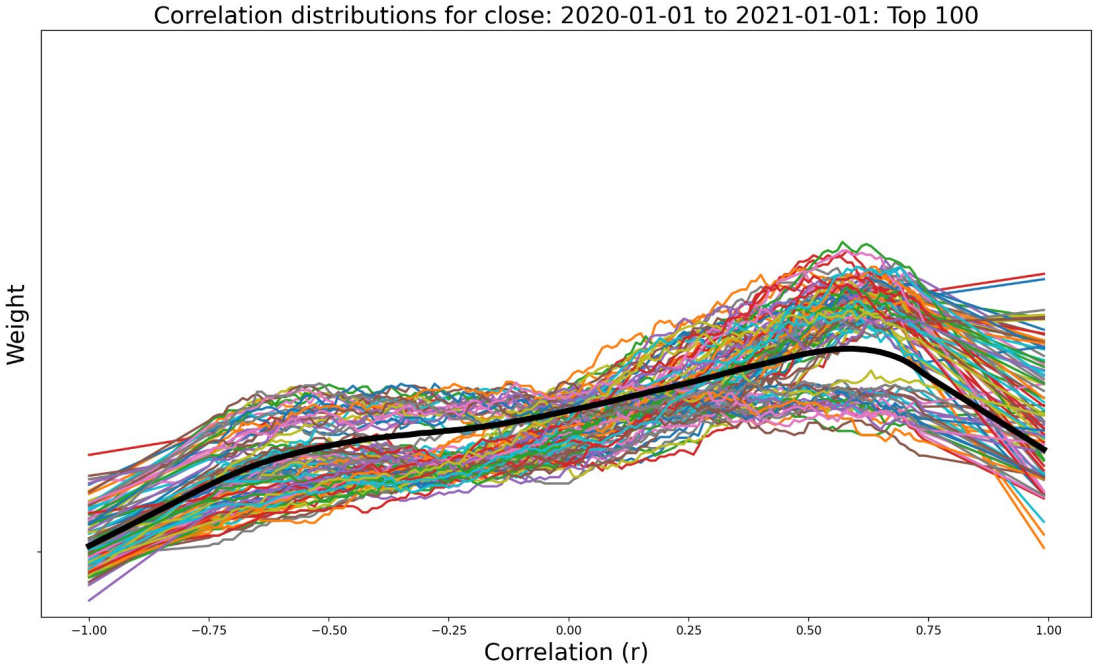
Results - 2019



There seem to be more outliers. Towel shows up once again, but I don't recognize anything else.

pos_score		neg_score	
BBBY	106.37	NEM	-84.79
JWN	98.23	LHX	-72.67
R	97.72	ASA	-72.19
KNX	90.05	RGLD	-71.84
GTES	88.98	CME	-70.72
LEA	88.22	MKTX	-68.44
NTB	87.71	HSY	-67.3
CPRI	87.41	HEI	-65.13
SAN	87.18	MDT	-64.68
WTFC	86.74	APD	-62.91
LFUS	86.38	SBUX	-62.23
IVZ	85.38	TSN	-62.01
COTY	85.06	PPC	-60.28
PENN	84.58	CLVT	-60.27
TCS	83.86	ATUS	-60.27
VFC	83.53	YUM	-60.26
SNA	83.38	MTCH	-59.94
NUE	83.26	SWCH	-59.69
WAL	82.64	ICE	-59.38
RGA	82.5	ROKU	-59.07

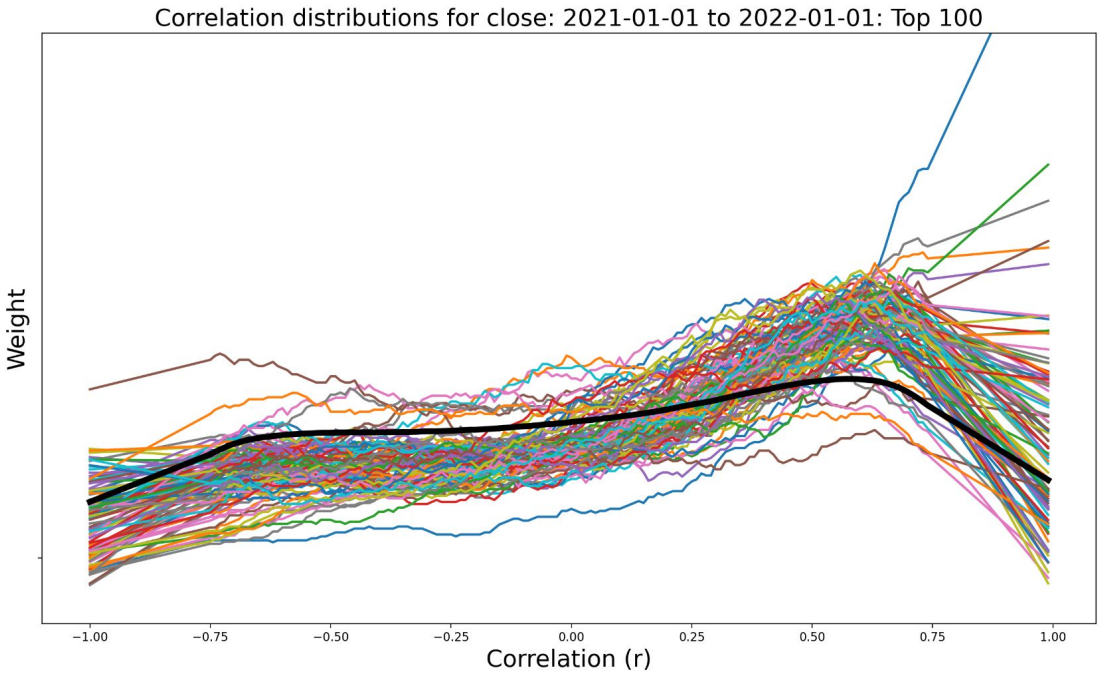
Results - 2020



Things look mostly normal-ish. I don't recognize any of the outliers, but this is still mostly before reddit crazines.

pos_score		neg_score	
WOLF	95.15	REGN	-57.69
AVT	94.07	MRNA	-57.19
CCL	93.63	NTWK	-51.23
OLED	92.66	HZNP	-48.7
JWN	91.37	HRL	-48.03
APH	91.33	QDEL	-47.74
ALK	90.96	PNW	-46.71
POST	90.32	HE	-46.44
OSW	90.31	NEM	-45.48
FDX	89.18	ATO	-45.24
GLW	89.06	LLY	-45.05
STLA	88.76	NVAX	-44.86
KSS	88.69	CMS	-44.57
NXPI	88.14	HOLX	-44.18
GPC	88.12	BDX	-43.69
KEYS	87.41	GO	-43.08
SABR	87.09	BAX	-43.03
HOG	87.04	D	-42.99
AA	86.99	TWLO	-42.89
WBD	85.47	CMPR	-42.65

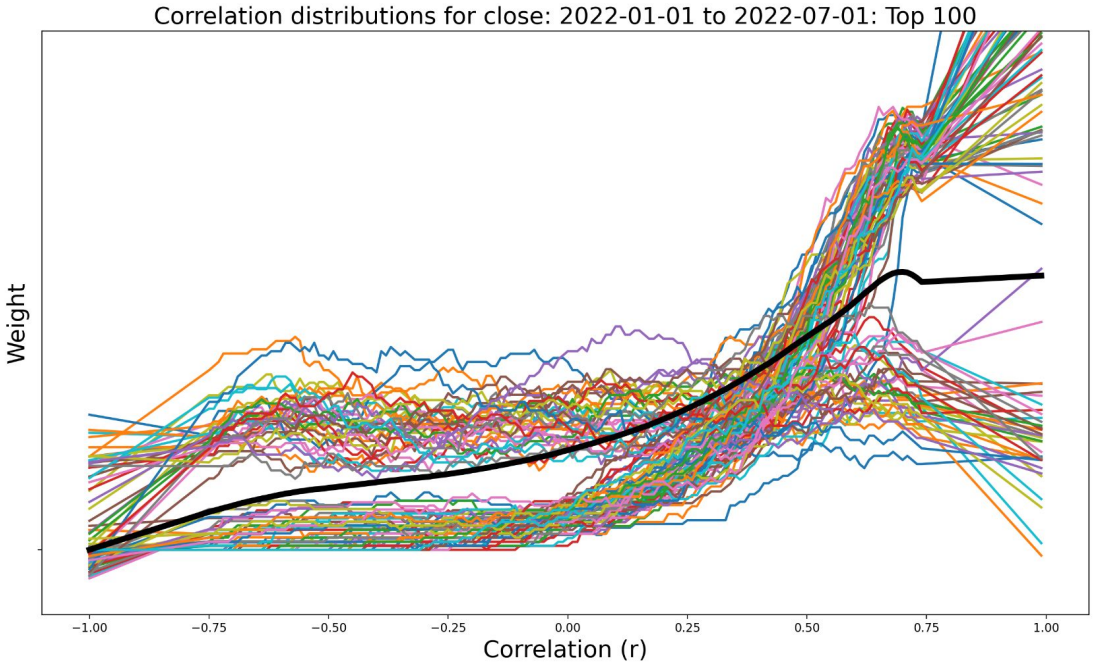
Results - 2021



Things have gone crazy. There a now 6+ clear outliers in the positive range, and wouldn't you know, it's a bunch of memestocks. This gives me some confidence in the validity of this method.

pos_score		neg_score	
AMC	157.36	AZO	-81.09
BB	114.94	BR	-78.37
BBBY	110.44	BKI	-72.68
PLUG	108.68	ZTS	-71.82
IRBT	100.12	ORLY	-71.42
PLTR	99.45	HSY	-69.02
BYND	96.26	AWK	-69
DKNG	90.21	MRK	-68.43
NCLH	89.34	VRSK	-67.78
EXAS	87.3	DUK	-67.64
W	86.48	ROL	-67.5
CCL	86	CPRT	-66.34
LESL	85.75	HUM	-66.03
WSM	85.48	CL	-65.75
CHPT	85.44	BIO	-65.56
AAL	85.01	PG	-65.3
QS	84.69	DG	-64.53
SKLZ	84.12	WAT	-64.41
CMBM	83.94	PEG	-63.93
ENOV	83.17	MASI	-62.9

Results - 2022

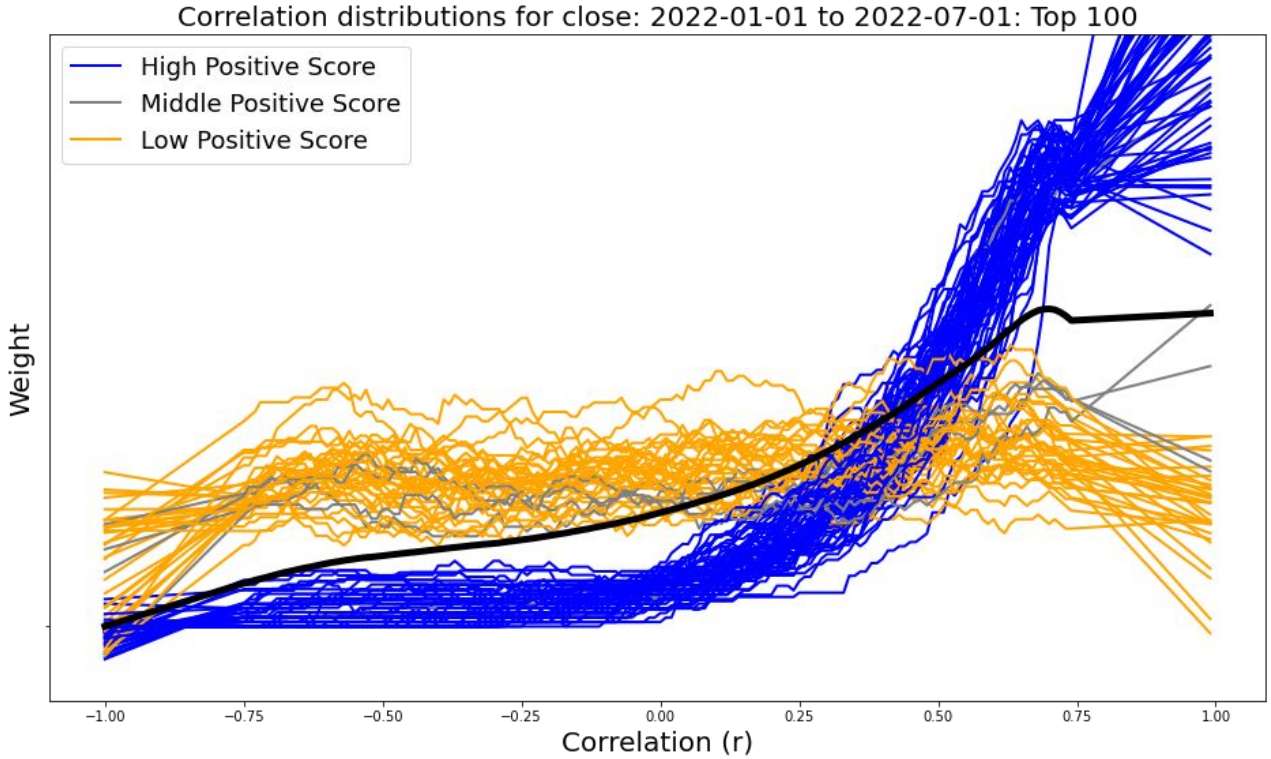


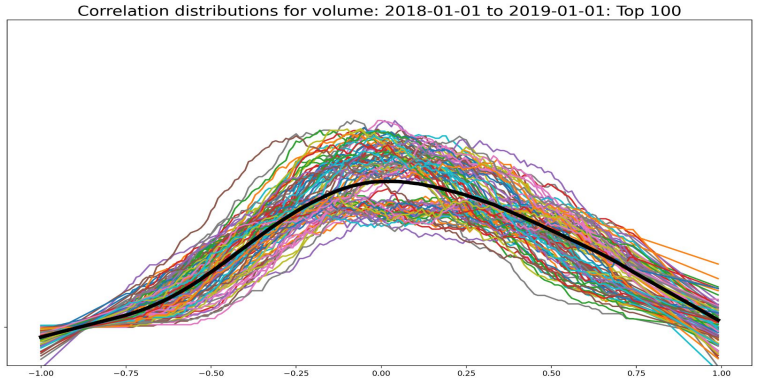
Something weird has happened. Why does this look so different from the other years? There also seem to be two different groups of stocks, those that follow the pattern of previous years, and those that are highly correlated with GME. A lot of stocks that were popular on reddit show up. Also note that the rolling scores are about half of the others because this covers only half a year, not a full year.

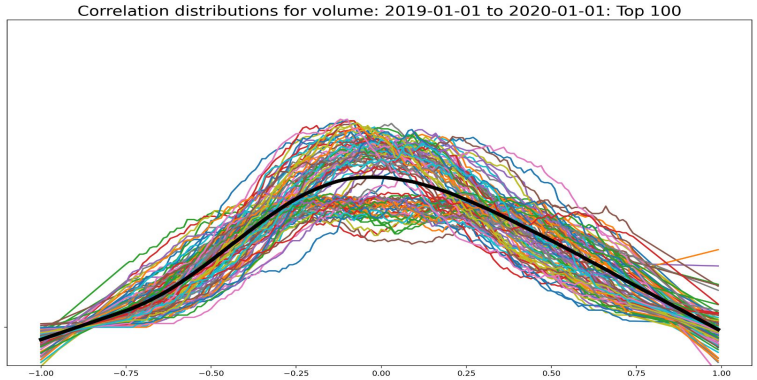
pos_score		neg_score	
AMC	99.18	XOM	-30.66
DOCU	85.45	MRK	-30.27
PLTR	84.84	K	-26.9
COTY	84.63	CLX	-25.6
BILL	84.32	SLVM	-24.66
SPCE	83.76	CVX	-23.9
RIVN	83.38	CPB	-23.25
ZS	82.95	XEL	-22.32
CHPT	82.65	KMB	-22.21
DNA	82.53	DINO	-22.09
TDOC	82.4	GIS	-22.09
COIN	82.2	ED	-21.94
PCOR	80.88	KHC	-21.64
DOCS	80.63	NRG	-21.53
ESTC	80.57	OXY	-21.53
HUBS	80.39	SJM	-21.29
ZM	80.12	KEX	-21.26
PLUG	79.91	HAL	-21.23
U	79.48	VRTX	-21.18
BSY	79.23	FLO	-20.42

Results - 2022 Grouping - Top 100

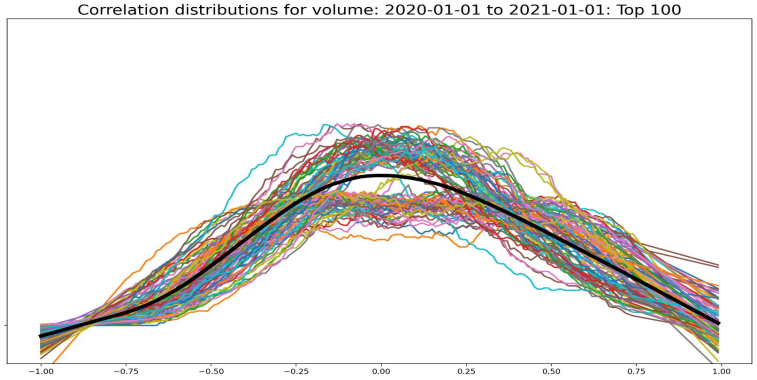
I wanted to take a closer look at what seems to be two distinct groups. These are the top 100 again, this time colored. Looks strange, but after coming back to this I think the orange is what the correlation of a random stock is supposed to look like, but there are so many blue stocks they warp the global distribution so badly that normal behavior is flagged as weird.

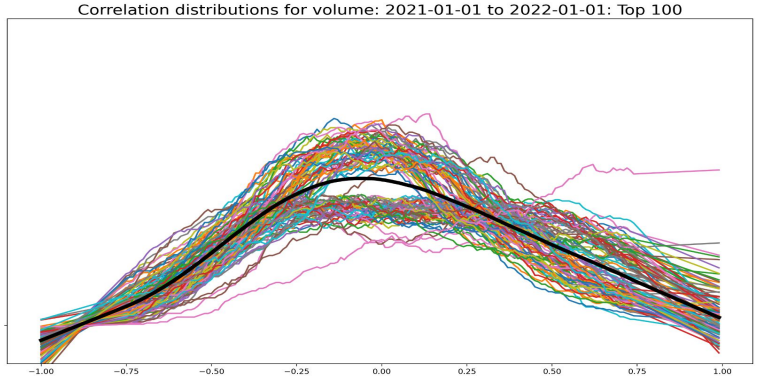






GME
Volume





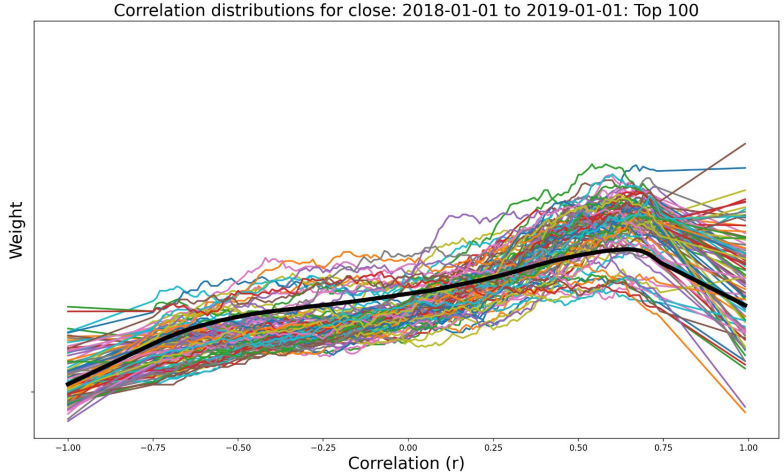
The same process but done using daily volume instead of closing price. Honestly, this is what I had expected to see. Something vaguely normally-distributed, with most stocks not being all that correlated. That is not what the closing price distributions show. In the 2021 chart (bottom right), the outlier stock is our friend, popcorn.

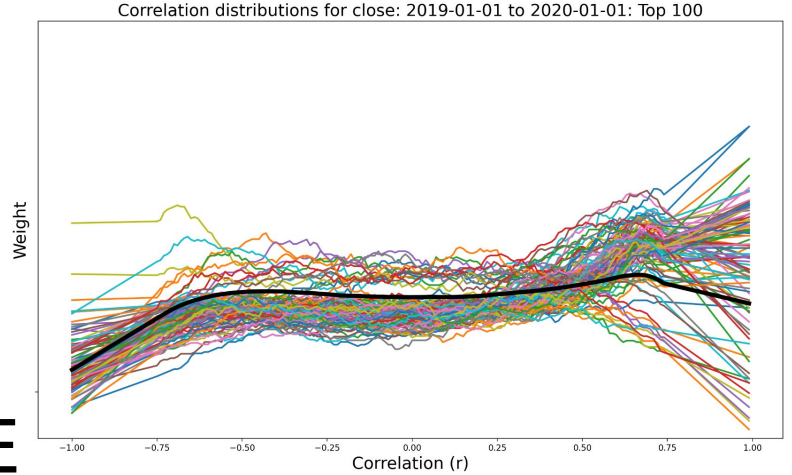
Results - Other Stocks

It’s hard to make any interpretations without looking at other stocks, so I ran the whole process with a couple of randomly picked ticker symbols:

- CHH - Choice Hotels International, Inc. - Consumer Cyclical
- INTU - Intuit - Technology
- RGLD - Royal Gold - Basic Materials
- WST - West Pharmaceutical Services, Inc. - Healthcare
- MSCI - MSCI, Inc. - Financial
- CCK - Crown Holdings, Inc. - Consumer Cyclical
- BBWI - Bath and Body Works, Inc. - Consumer Cyclical
- LECO - Lincoln Electric Holdings, Inc. - Industiral
- EVRG - Evergy Inc - Utilities

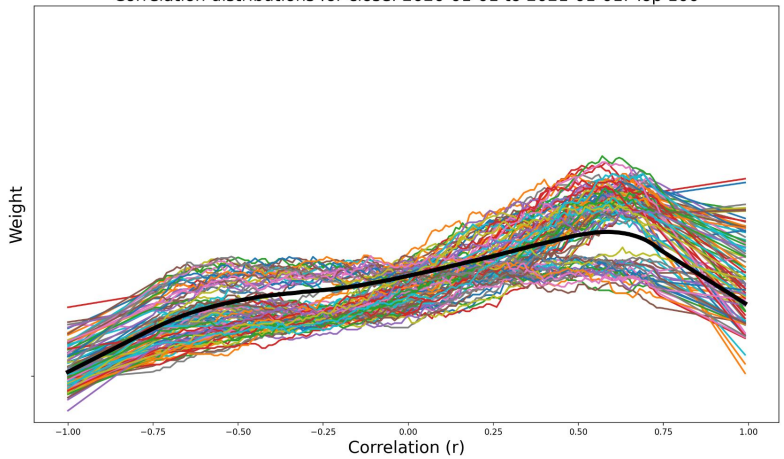
For easy comparison, I’ve included GME again in the same format

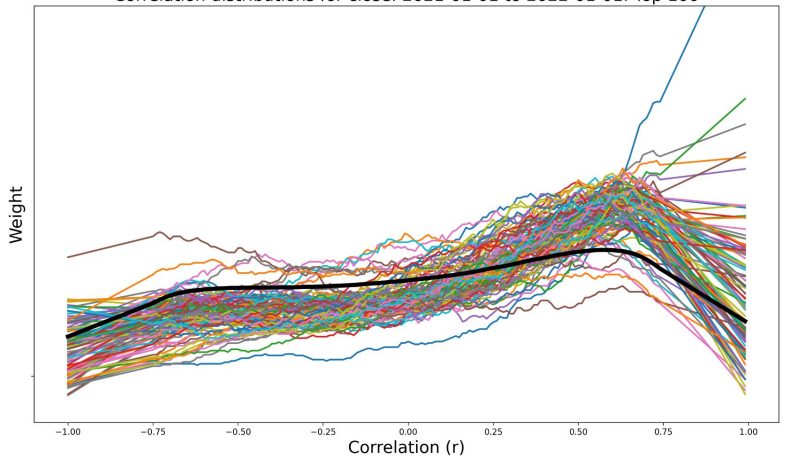


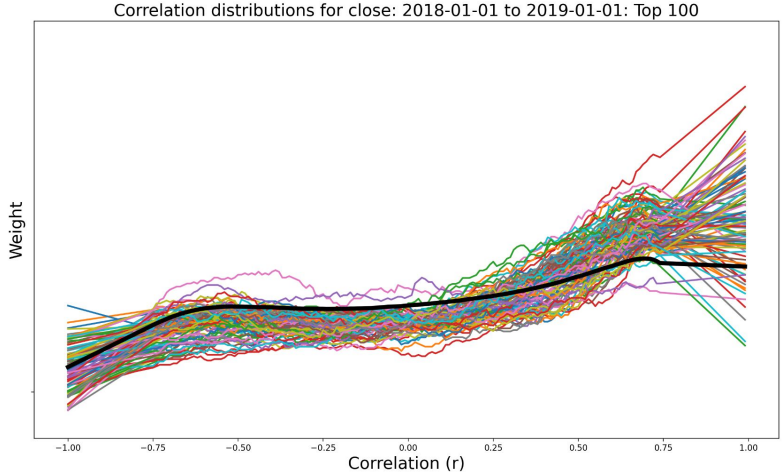


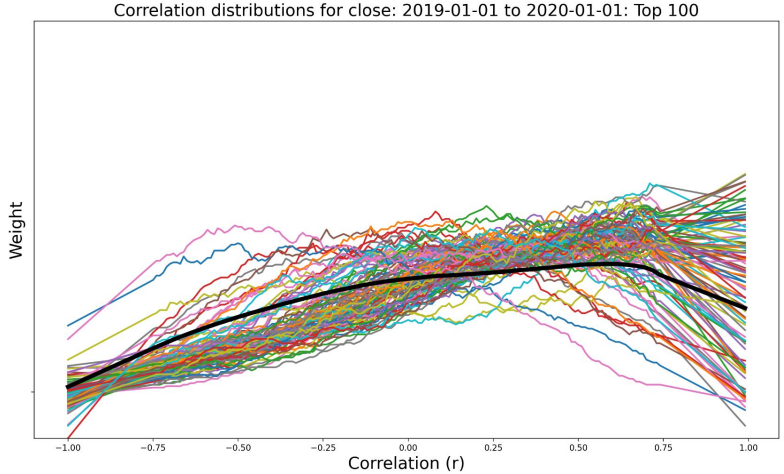
GME

Closing price

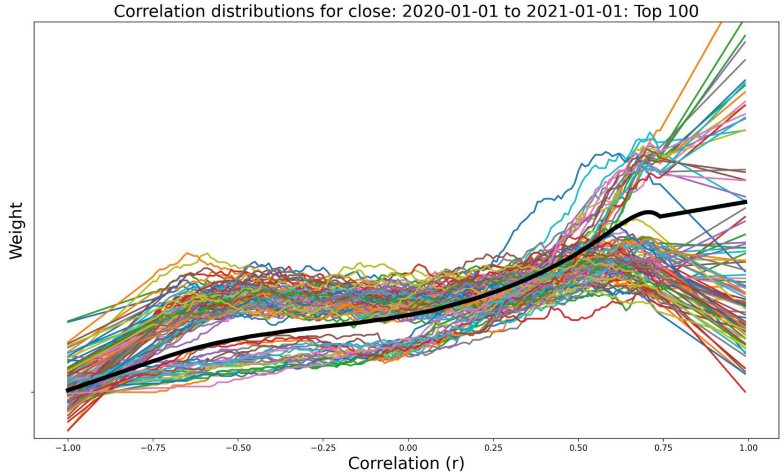


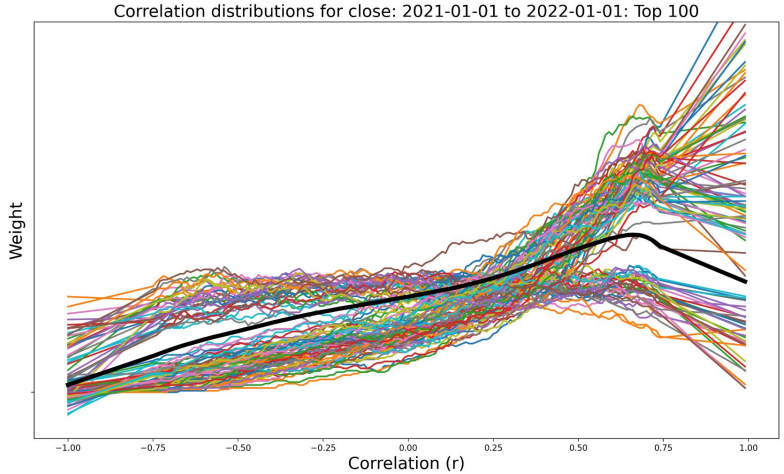


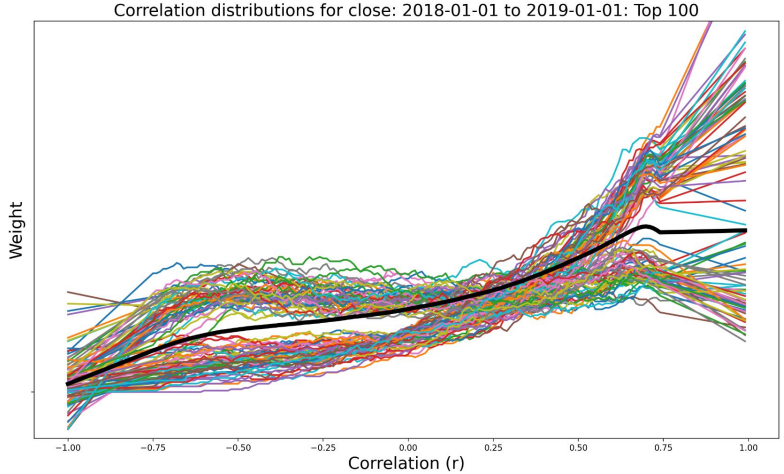

Correlation distributions for close: 2018-01-01 to 2019-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

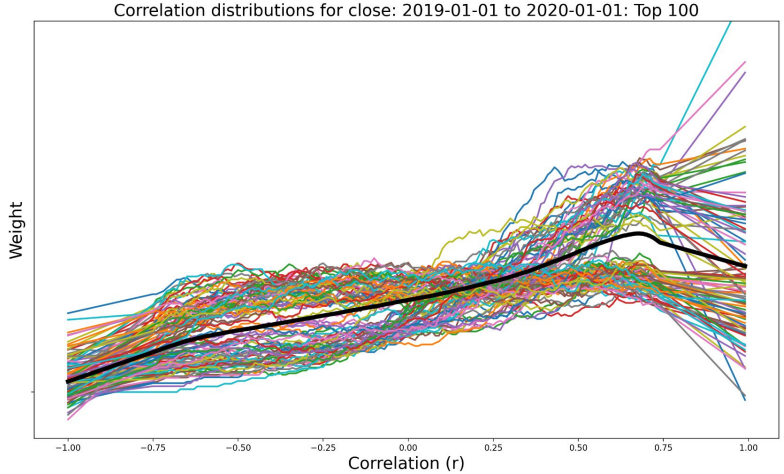

Correlation distributions for close: 2019-01-01 to 2020-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

BBWI

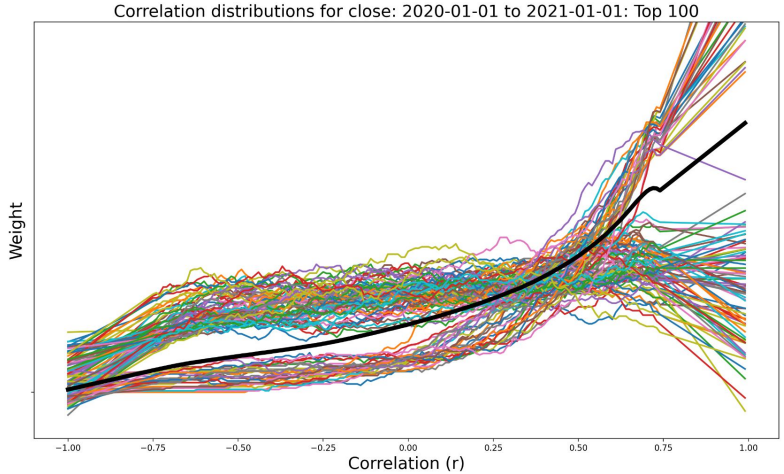

Correlation distributions for close: 2020-01-01 to 2021-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

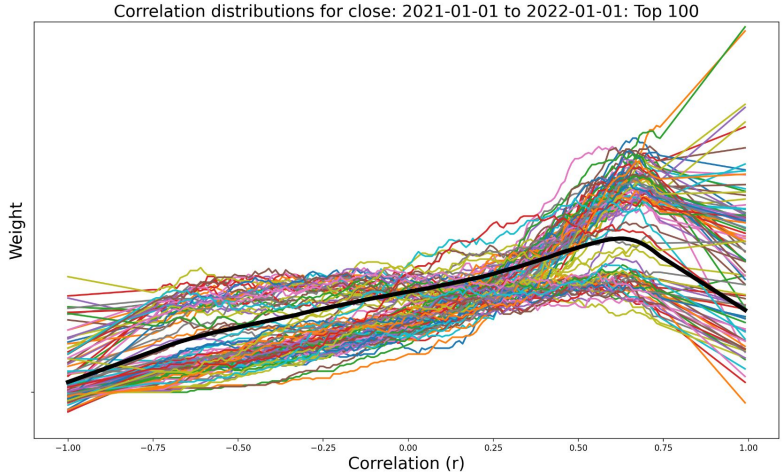

Correlation distributions for close: 2021-01-01 to 2022-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

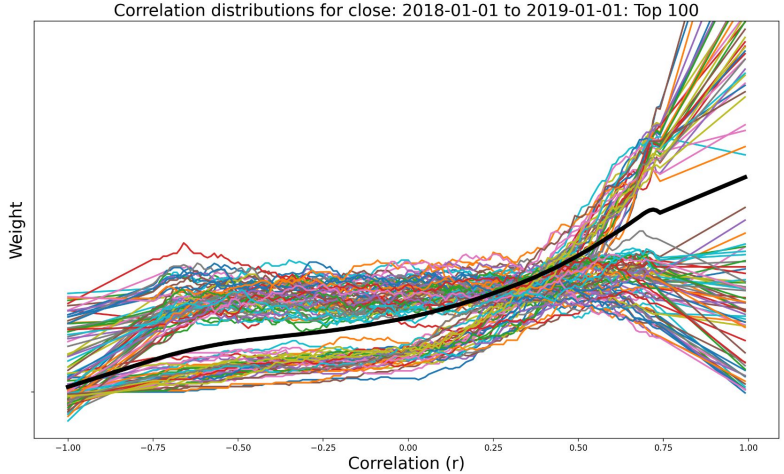

Correlation distributions for close: 2018-01-01 to 2019-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

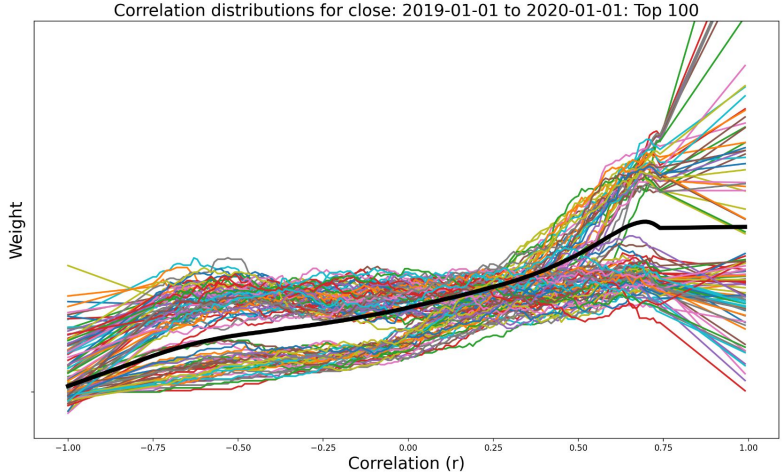

Correlation distributions for close: 2019-01-01 to 2020-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

CCK

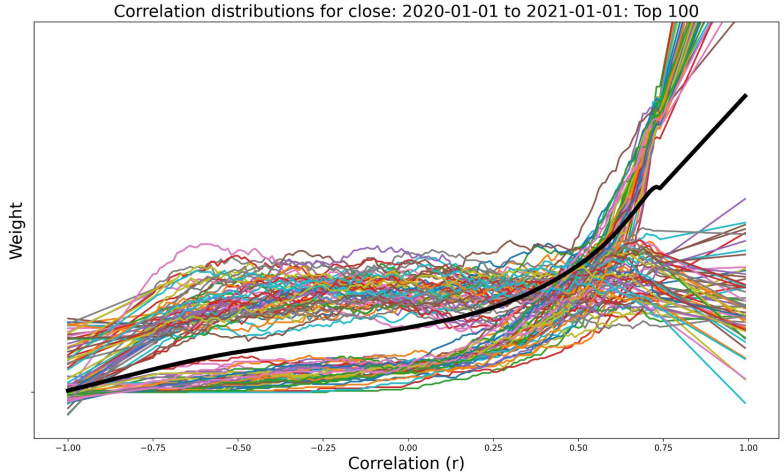

Correlation distributions for close: 2020-01-01 to 2021-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

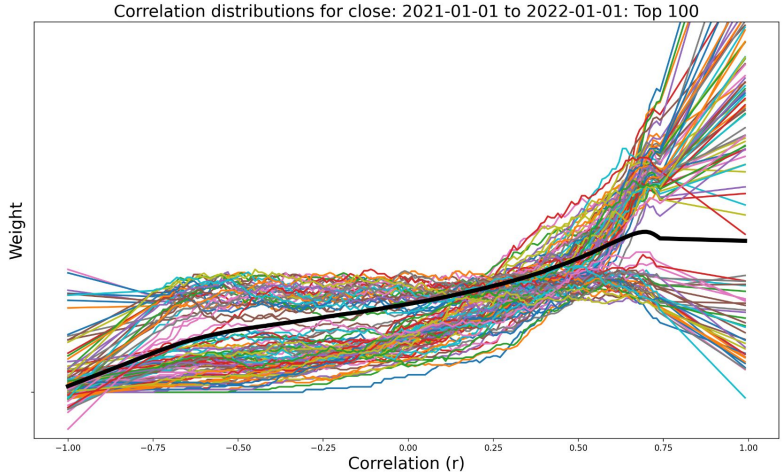

Correlation distributions for close: 2021-01-01 to 2022-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

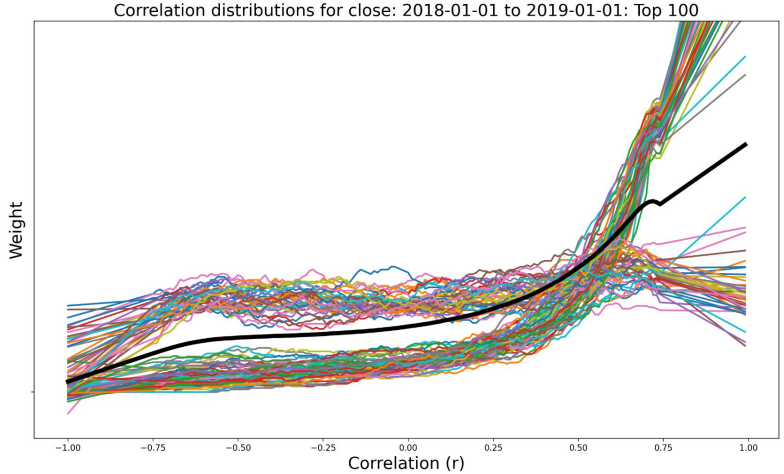

Correlation distributions for close: 2018-01-01 to 2019-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

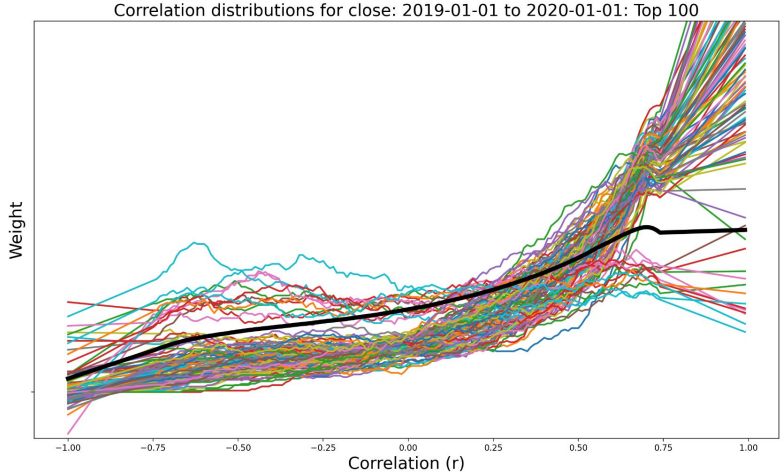

Correlation distributions for close: 2019-01-01 to 2020-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

CHH

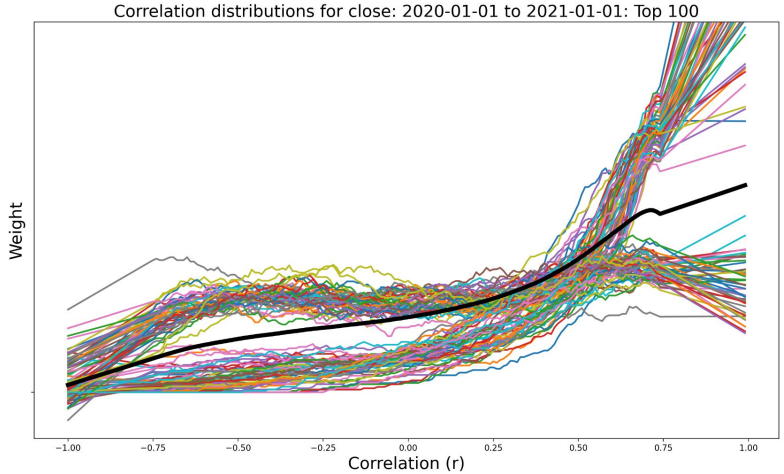

Correlation distributions for close: 2020-01-01 to 2021-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

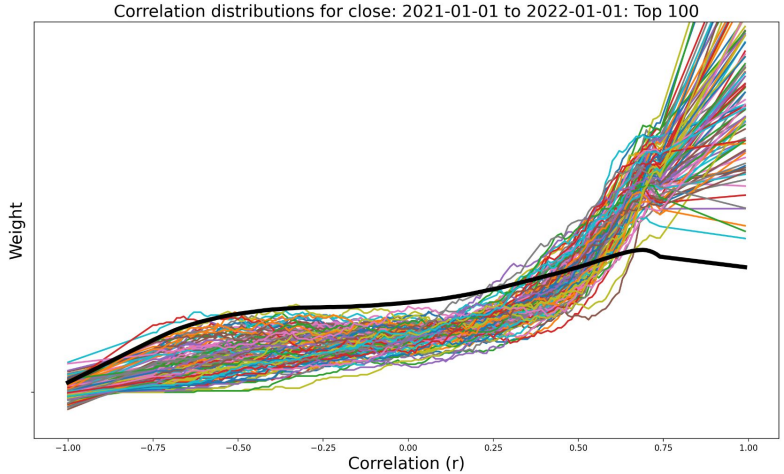

Correlation distributions for close: 2021-01-01 to 2022-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

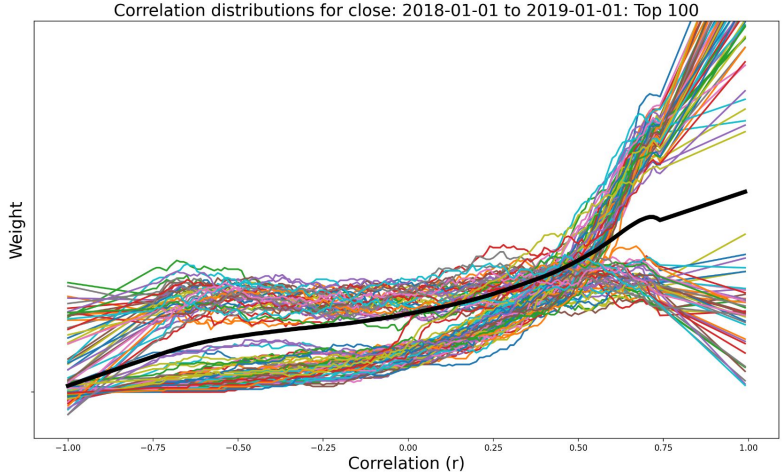

Correlation distributions for close: 2018-01-01 to 2019-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

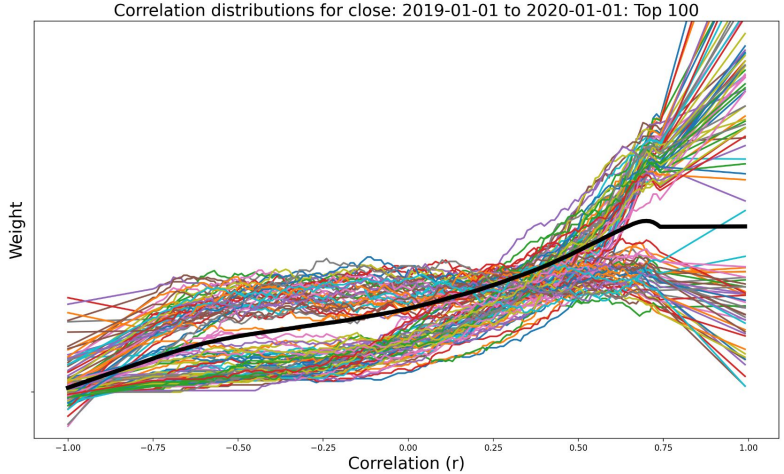

Correlation distributions for close: 2019-01-01 to 2020-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

INTU

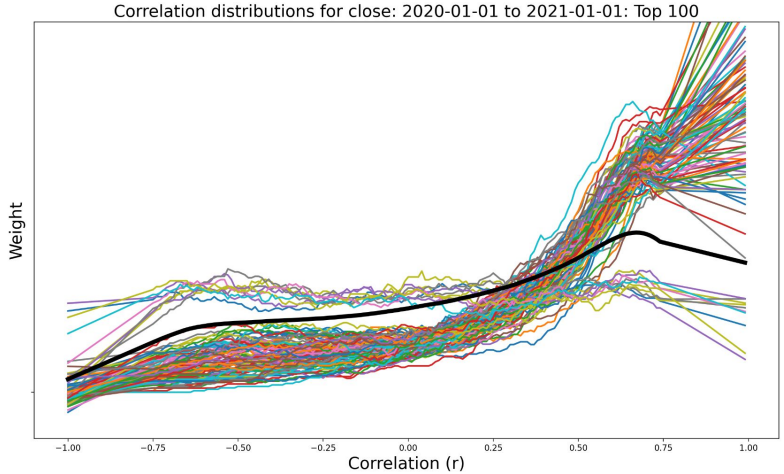

Correlation distributions for close: 2020-01-01 to 2021-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

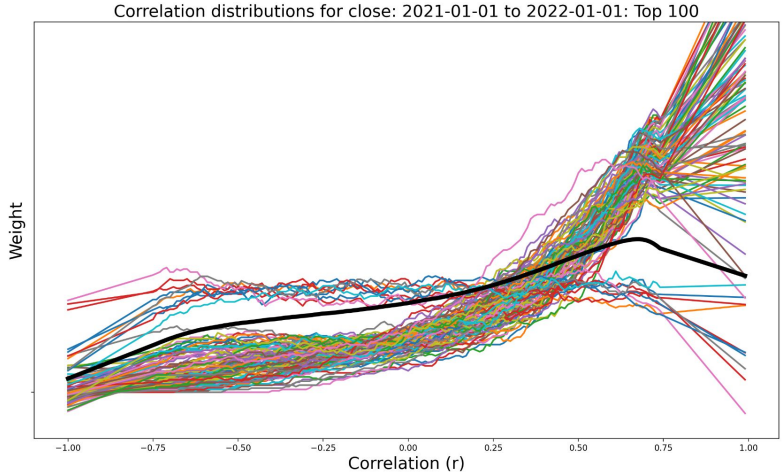

Correlation distributions for close: 2021-01-01 to 2022-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

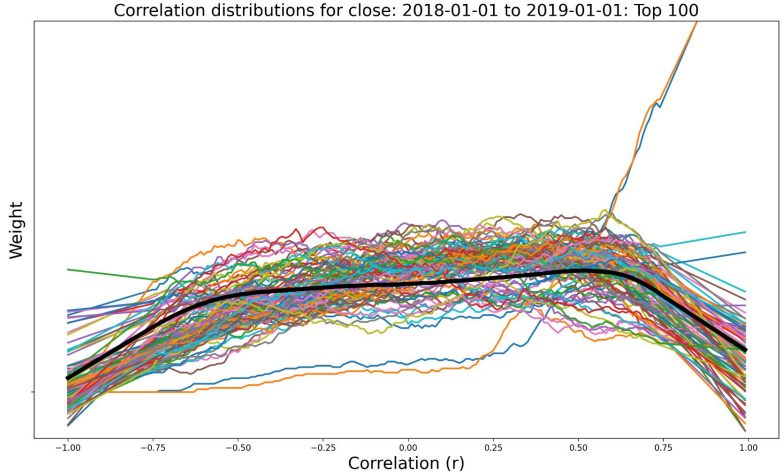

Correlation distributions for close: 2018-01-01 to 2019-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

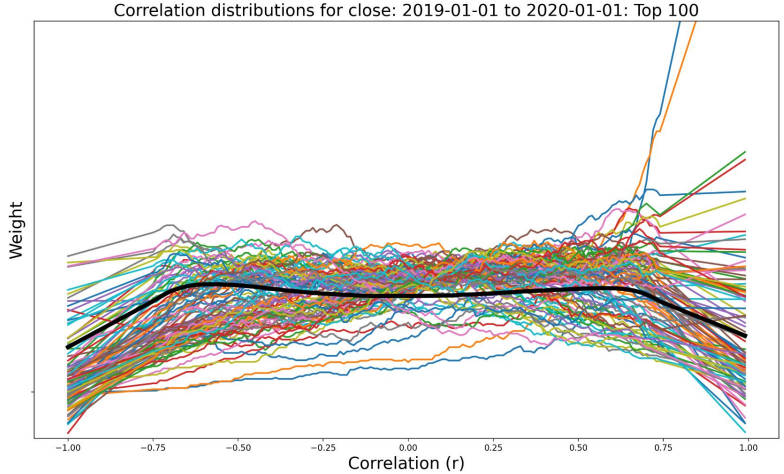

Correlation distributions for close: 2019-01-01 to 2020-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

MSCI

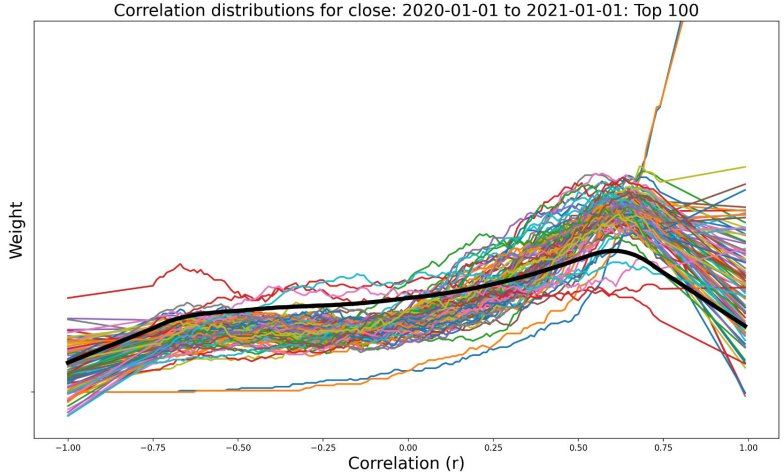

Correlation distributions for close: 2020-01-01 to 2021-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

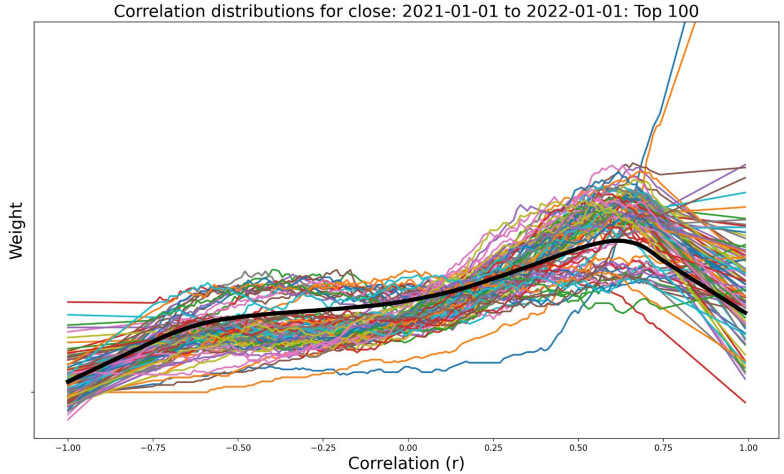

Correlation distributions for close: 2021-01-01 to 2022-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

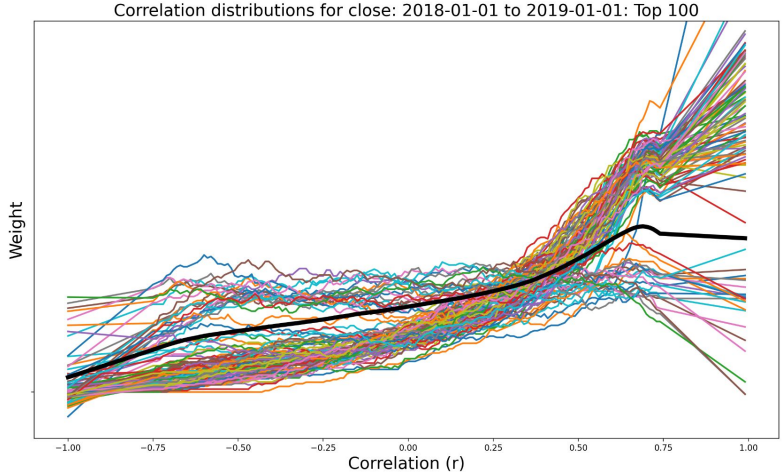

Correlation distributions for close: 2018-01-01 to 2019-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

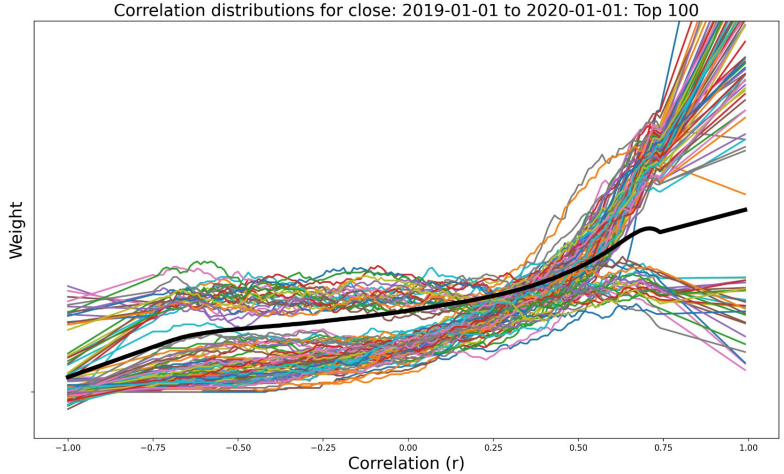

Correlation distributions for close: 2019-01-01 to 2020-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

RGLD

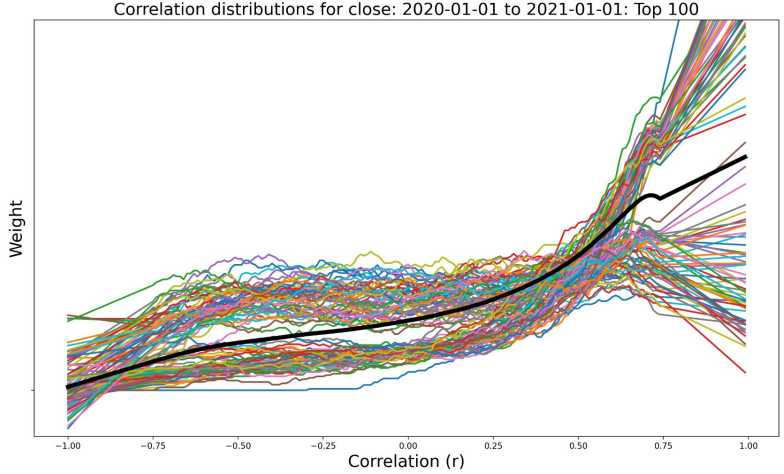

Correlation distributions for close: 2020-01-01 to 2021-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

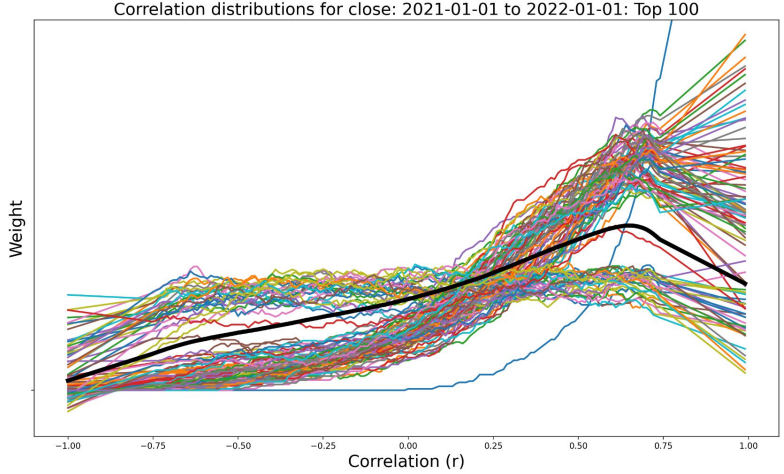

Correlation distributions for close: 2021-01-01 to 2022-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

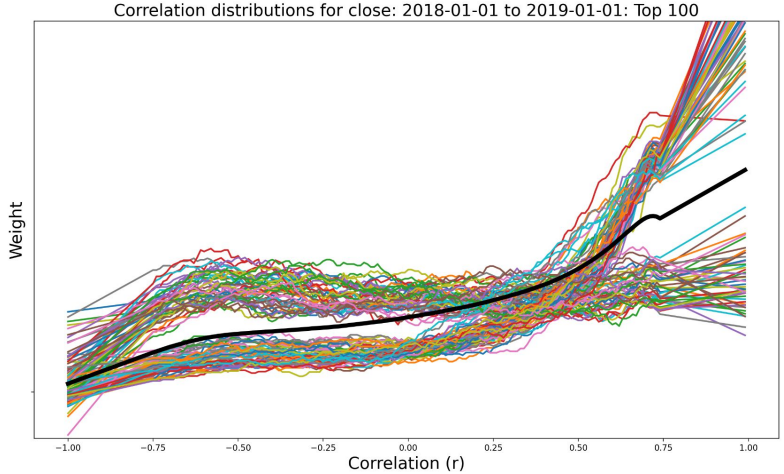

Correlation distributions for close: 2018-01-01 to 2019-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

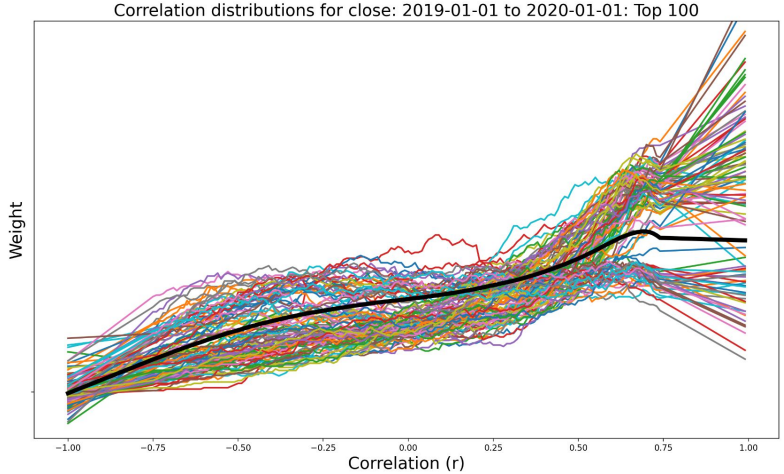

Correlation distributions for close: 2019-01-01 to 2020-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

SCCO

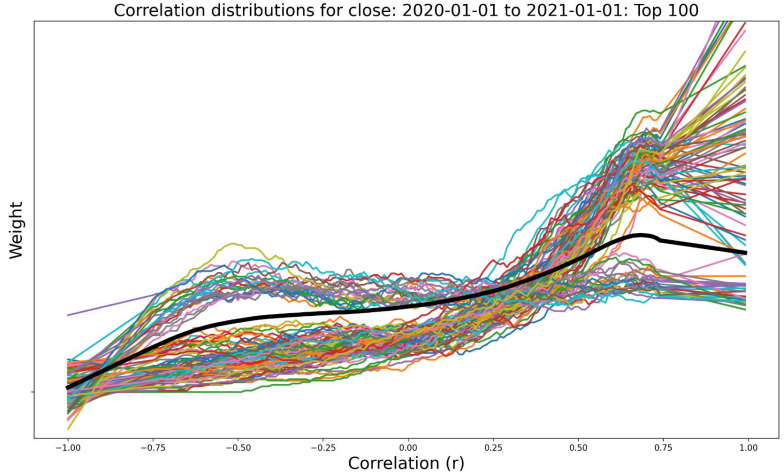

Correlation distributions for close: 2020-01-01 to 2021-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

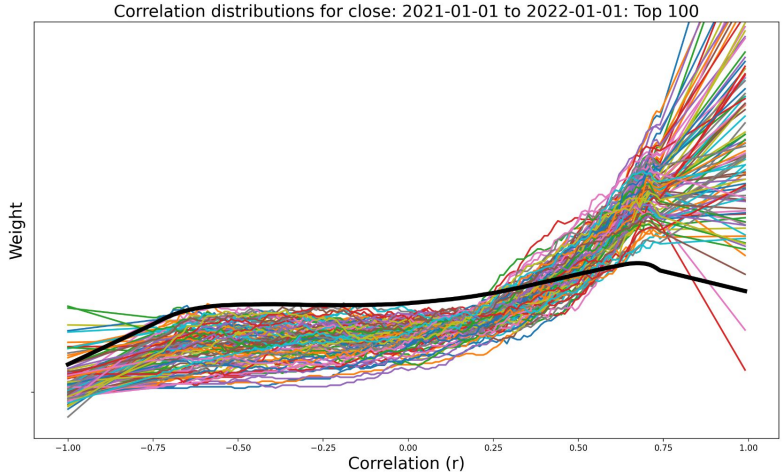

Correlation distributions for close: 2021-01-01 to 2022-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

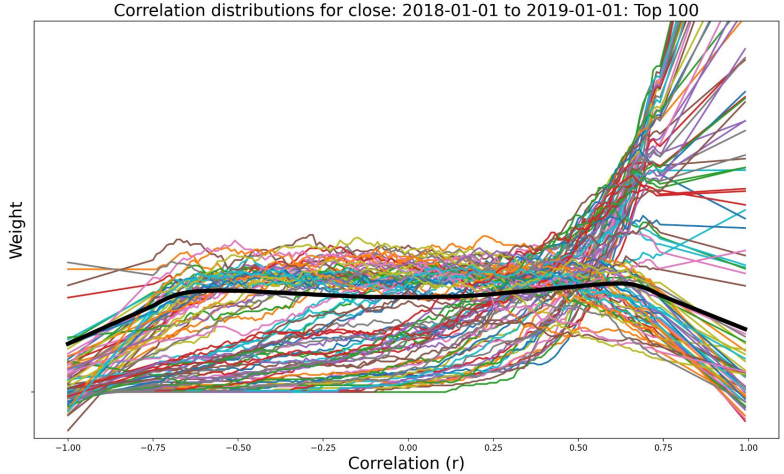

Correlation distributions for close: 2018-01-01 to 2019-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

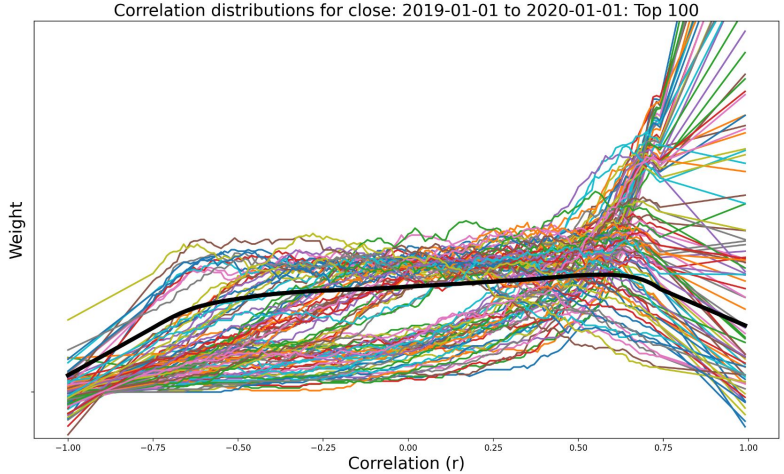

Correlation distributions for close: 2019-01-01 to 2020-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

WST

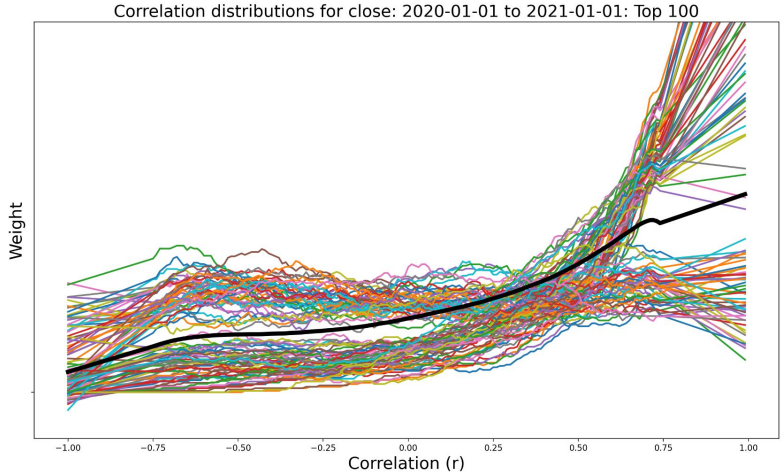

Correlation distributions for close: 2020-01-01 to 2021-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

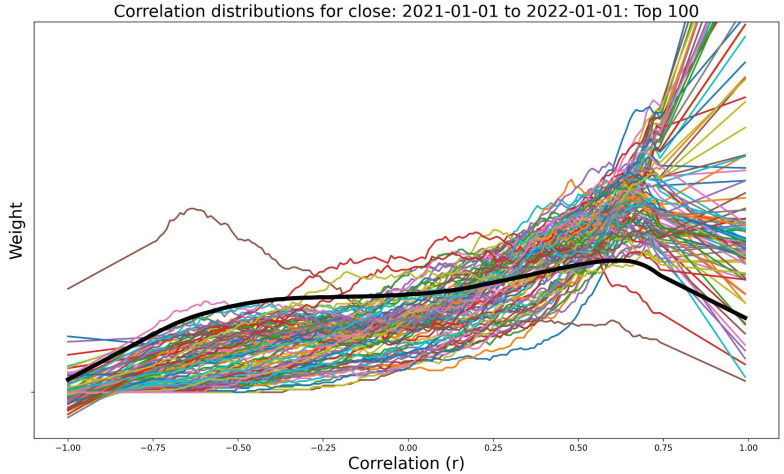

Correlation distributions for close: 2021-01-01 to 2022-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

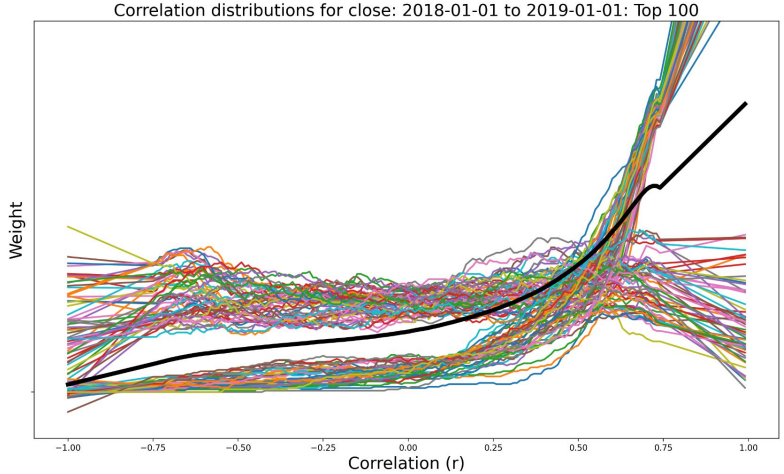

Correlation distributions for close: 2018-01-01 to 2019-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

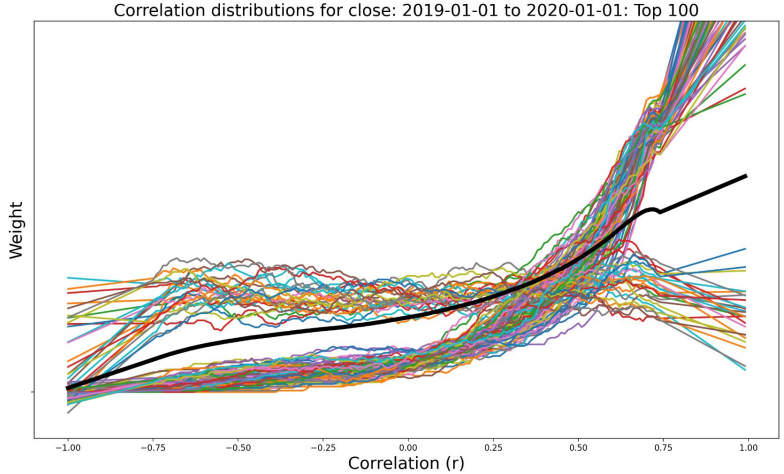

Correlation distributions for close: 2019-01-01 to 2020-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

EVRG

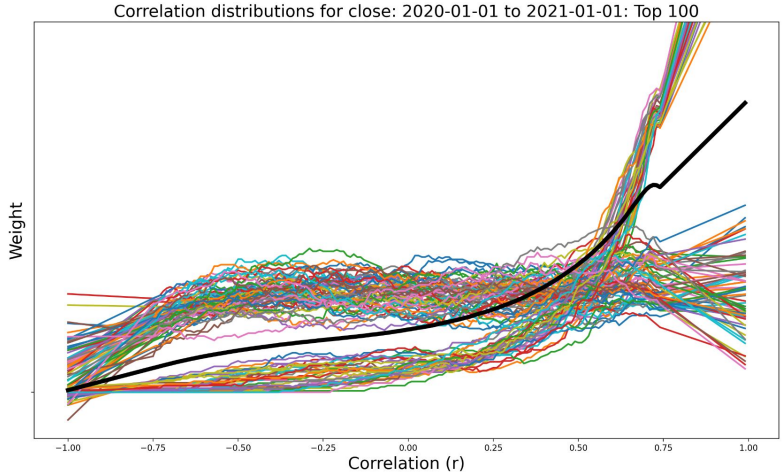

Correlation distributions for close: 2020-01-01 to 2021-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

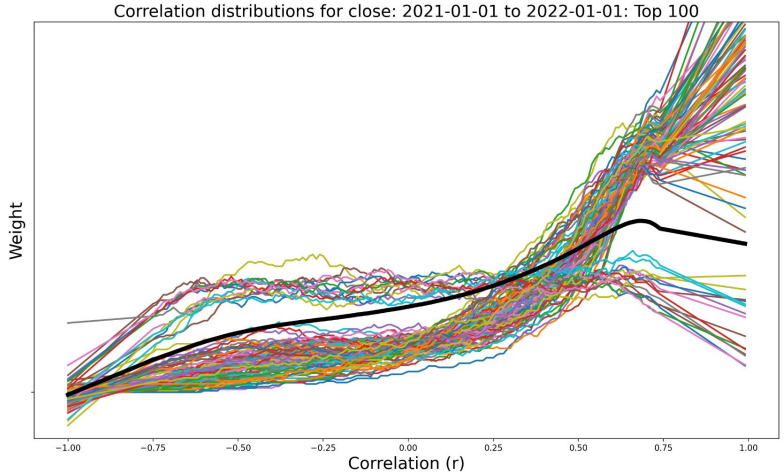

Correlation distributions for close: 2021-01-01 to 2022-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)


Correlation distributions for close: 2018-01-01 to 2019-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)


Correlation distributions for close: 2019-01-01 to 2020-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

LECO


Correlation distributions for close: 2020-01-01 to 2021-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)


Correlation distributions for close: 2021-01-01 to 2022-01-01: Top 100
Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation (r)

Discussion

I have a couple thoughts after looking at this random assortment of other stocks:

1. I expected something like the GME volume plots, or at least what RGLD shows, where the global line is flat, with maybe a few outliers.
2. A lot of them look like GME in the first half of 2022. Is this normal or are those stocks also behaving strangely? On the one hand, they sure do look weird, but on the other I picked them at random. It would be a huge coincidence if that many stocks were unusually correlated with each other. Perhaps this *is* normal behavior, i.e. it's normal for stocks to lump up in correlations due to market forces, algorithmic behavior, etc. This means that the positive rolling score could still be useful.
3. I think that the two groups (blue and orange plots) that I was seeing was a result of having many stocks being strongly correlated, warping the global distribution so much that regular distributions are showing up as strange. It might be better to create the global distribution from the correlation between all possible pairs of stocks, over the whole time period, then use that as the global distribution to compare correlation distributions against.

Conclusions

Well, I've concluded that I still don't really know anything. If anything, I have more questions than when I started, but maybe reddit can help.

1. On the topic of the method:
    a. It seems like it could be valid, using closing price it captured many, if not all, the memestocks that exploded during 2021
    b. Volume behaves like I expected independent and uncorrelated stocks to behave, but it did not capture any of the weirdness that closing price did
2. On the topic of positive rolling scores:
    a. Is it really suspicious to have a high score here? I think so, but I'm not sure how to prove it.
    b. What about high scoring stocks that were not part of the memestock craze? Do they have anything that ties them to GME?
3. On the topic of distributions:
    a. Why do the other stocks have tons of stocks with high correlation? Is this normal? Are they included in the same ETFs, baskets, industries, hedge fund algos, etc.?

How you can help

Clone/fork the project and run it for yourself. If you have ideas/criticisms please let me know. Especially if I've made a mistake somewhere.

If you know how to research rabbit-holes and connections between stocks, try looking at the stocks that have high positive rolling scores with GME, but weren't involved with reddit or memestocks. Maybe this can be used as a way to guide searches for other suspicious stocks besides popcorn.