Stocks Clustering and Time Series Algorithms

Gaukhar Javarova

AGENDA

- Data introduction
- Technical Indicators
- Feature Engineering
- K-Means Clustering
- PCA, UMAP, TSNE
- LABELS (silhouette, plots, math)
- TIME SERIES ANALYSIS

Data Introduction

Highly fluctuated sectors of the market due to covid19

Oil

Precious metals

Airlines

Marine transportation

Auto Manufacturing

Biotech companies working on covid cure

Conferencing Platforms

Technical Indicators

RSI - Relative strength index, rsi = 70 overbought, rsi = 30 oversold.

SMA - Simple moving average. 200 MA and 50 MA - crosses up - goes up and vise versa.

EMA - The exponential moving average, is more sensitive to the price movements, more weight to recent price data.

Midpoint = Average (Highest Close - Lowest Close) within the look back period.

CCI - The commodity channel index (CCI) is an oscillator indicator that helps show when an asset has been overbought or oversold.

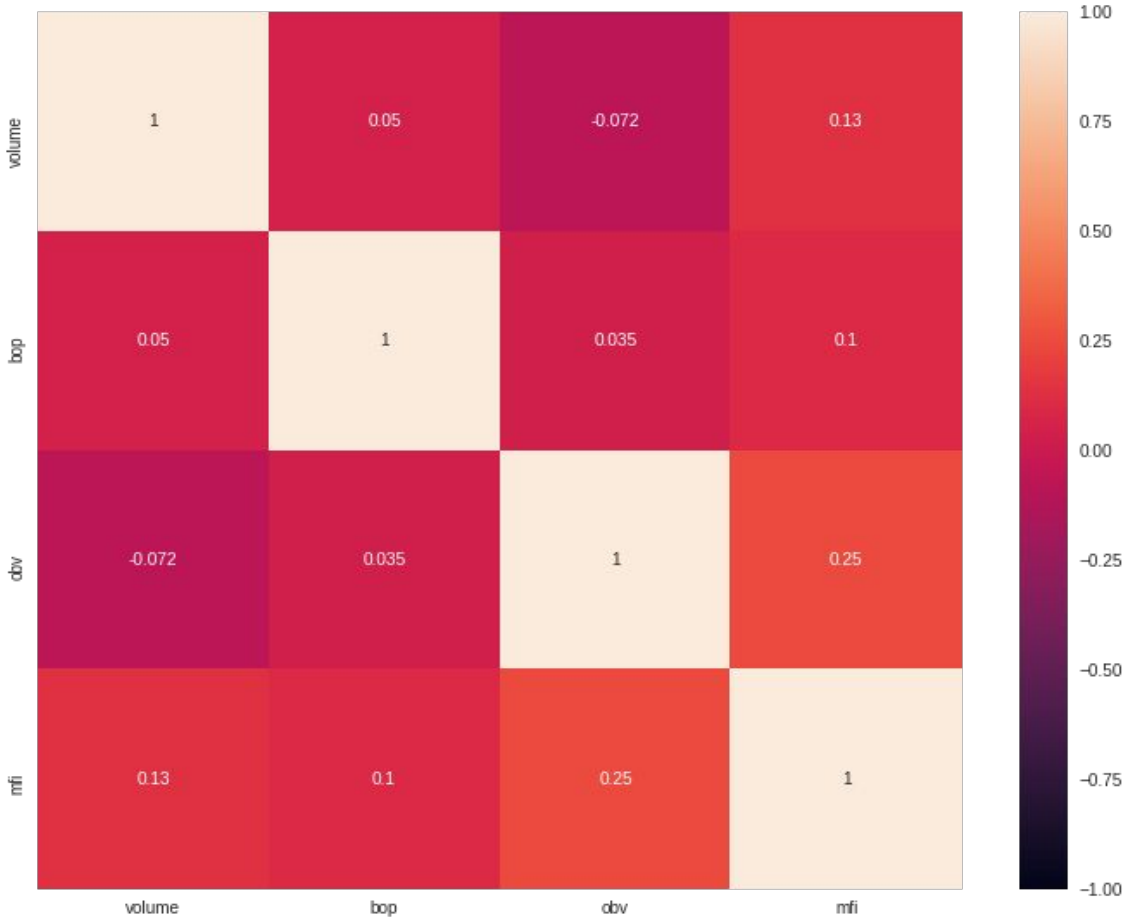
OBV - On-balance volume is a technical trading momentum indicator that uses volume flow to predict changes in stock price.

BOP - Balance of Power indicator measures the market strength of buyers against sellers. Potential trend reversal.

MFI - The Money Flow Index is a technical oscillator that uses price and volume data for identifying overbought or oversold signals in an asset, moves between 0 and 100.

Midpoint = Average (Highest Close - Lowest Close) within the look back period.

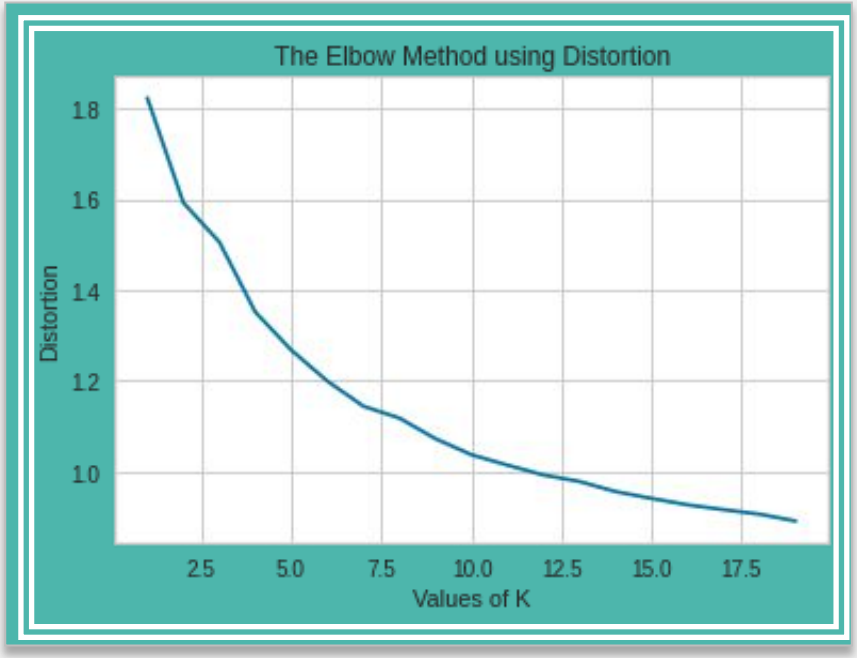
Feature Engineering


volume
bop
obv
mfi
1
0.05
-0.072
0.13
0.05
1
0.035
0.1
-0.072
0.035
1
0.25
0.13
0.1
0.25
1
volume
bop
obv
mfi
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Clean data

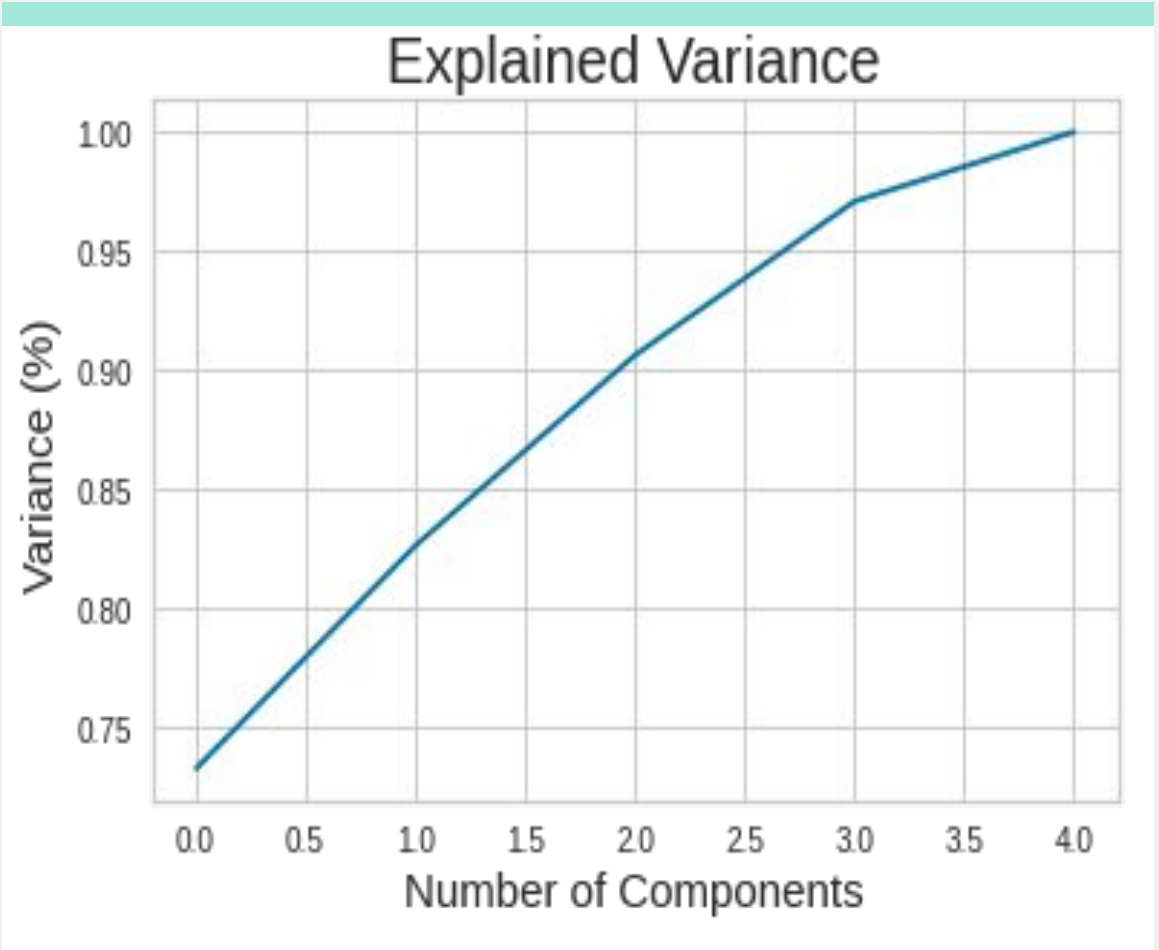
K-Means Clustering

The Elbow method to choose the number of clusters: 5 -10



PCA

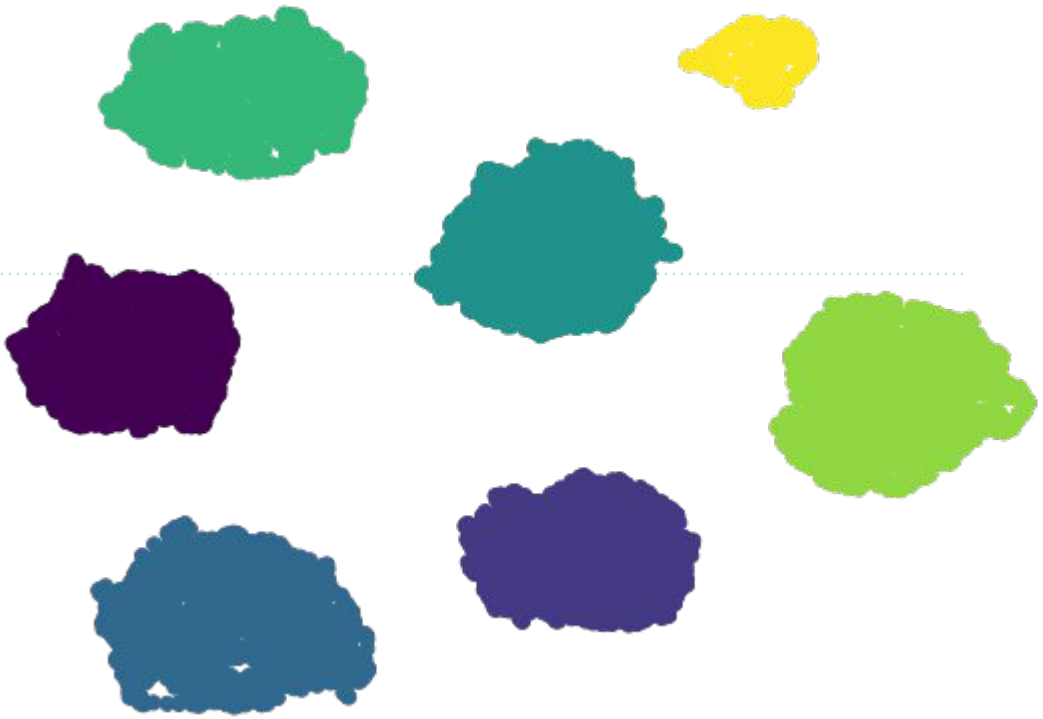
98% explained variance ratio for
first 3.0 components



PCA

silhouette score

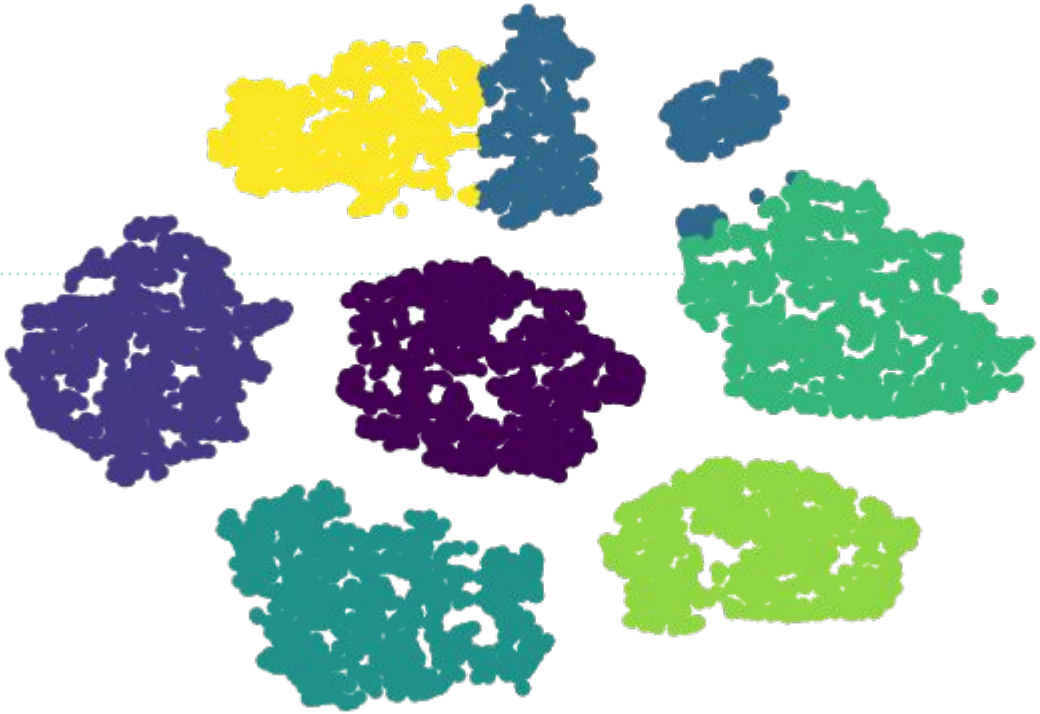
0.42902192731649114



UMAP

silhouette score

0.44640164689272954

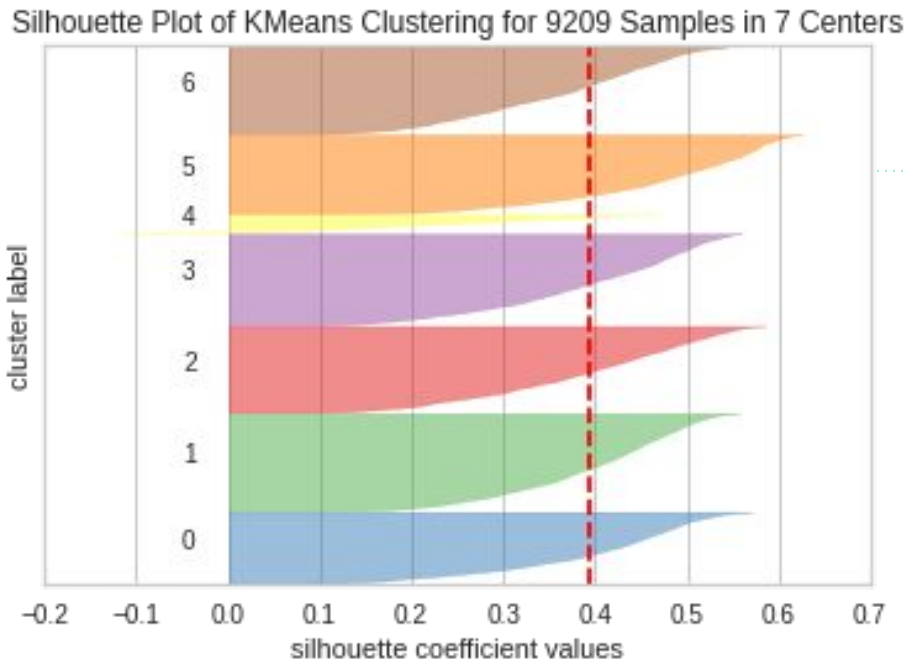


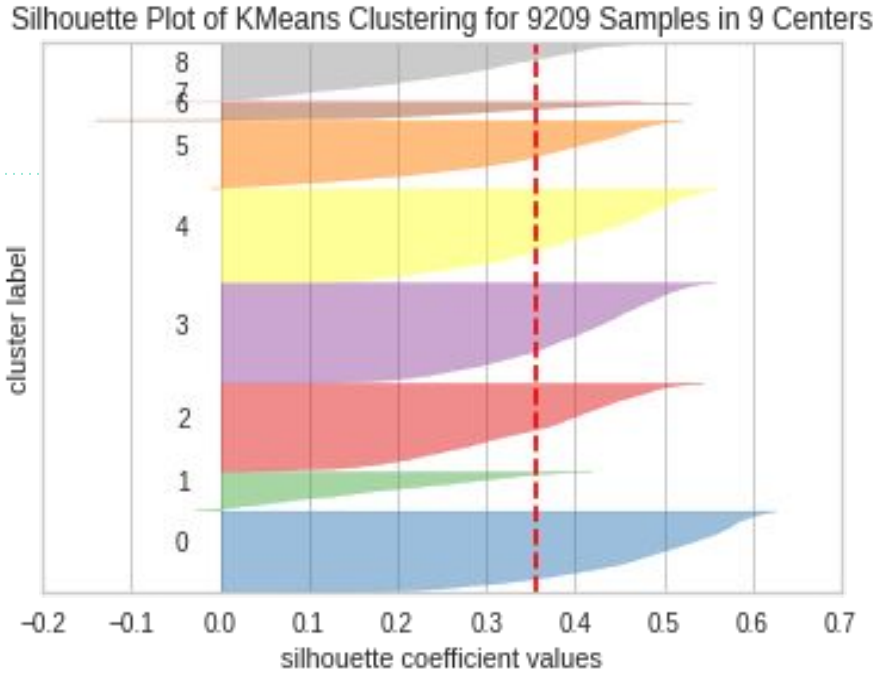
TSNE

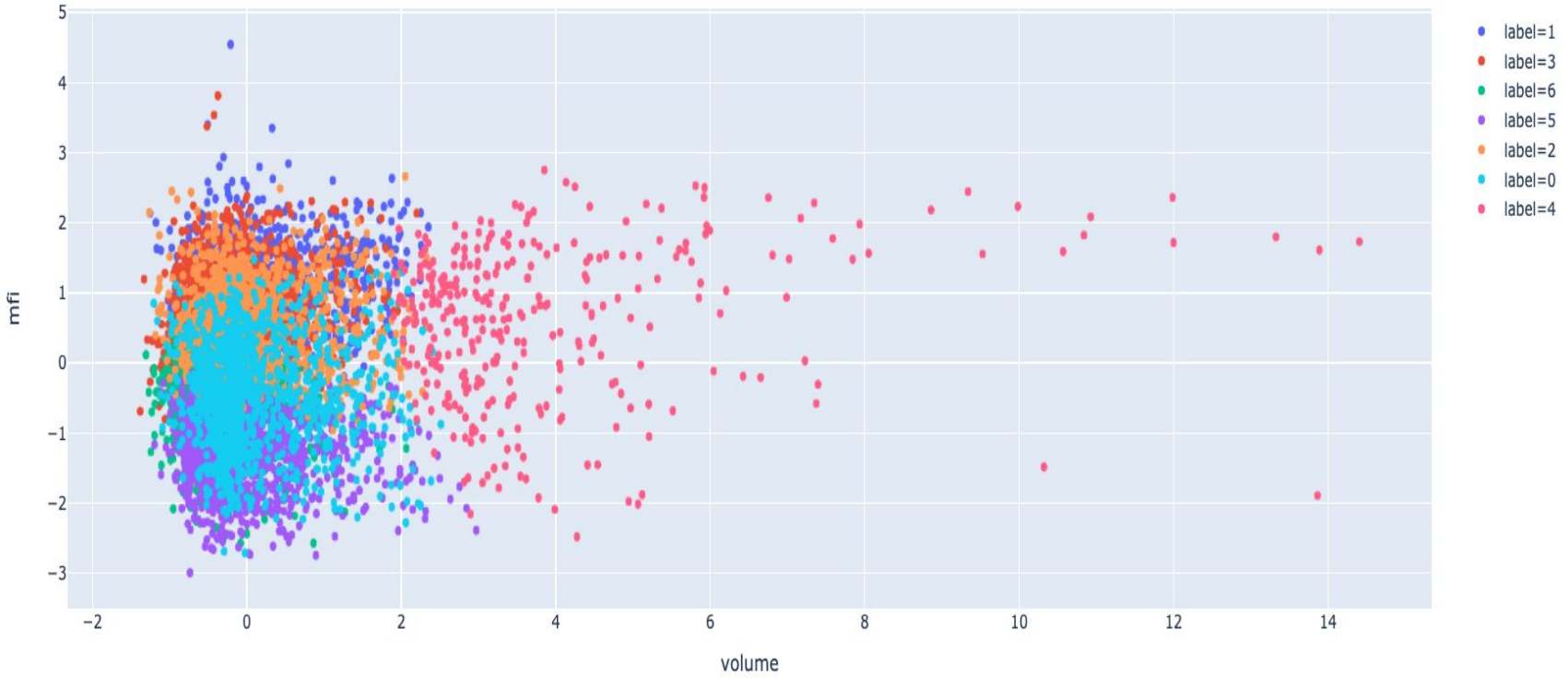
silhouette score

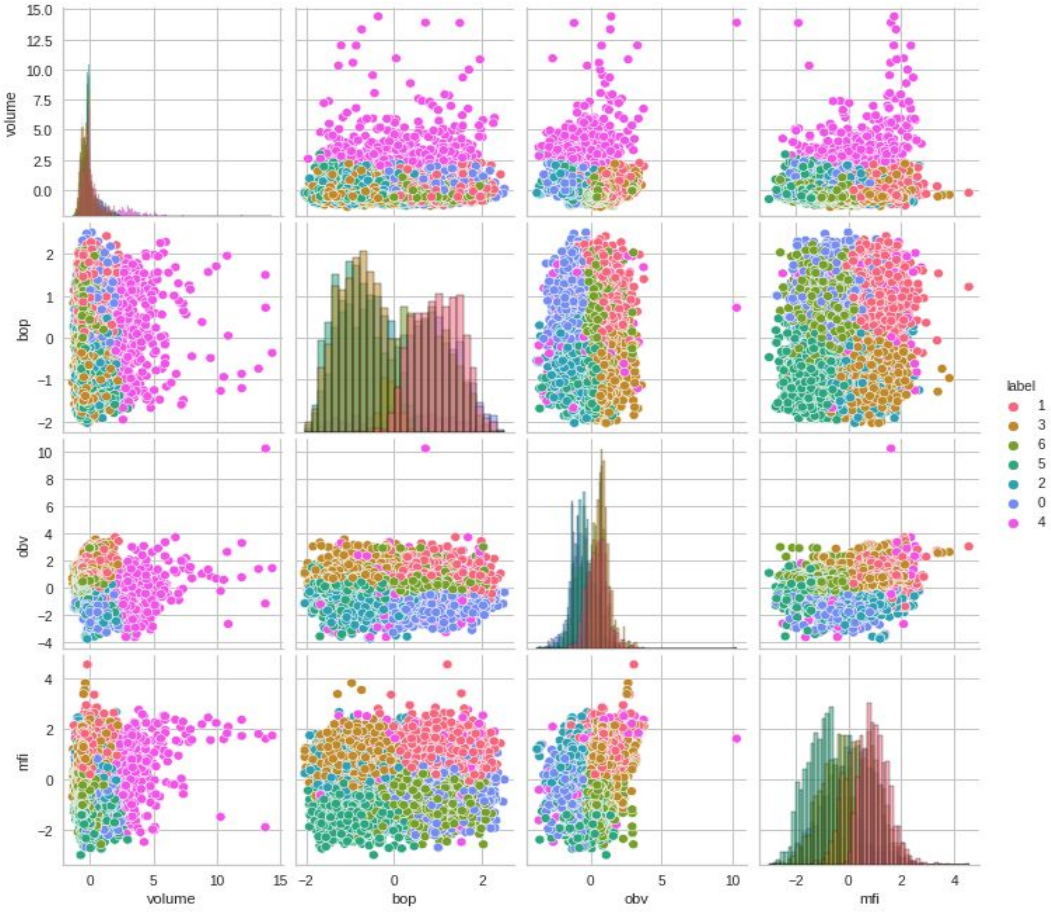
0.3372904014865667

Silhouette Visualizer






mfi
volume
5
4
3
2
1
0
−1
−2
−3
−2
0
2
4
6
8
10
12
14
label=1
label=3
label=6
label=5
label=2
label=0
label=4



Cluster 4 has the most volume - can go long and short

Cluster 5 is low on MFI (rsi) - can go long

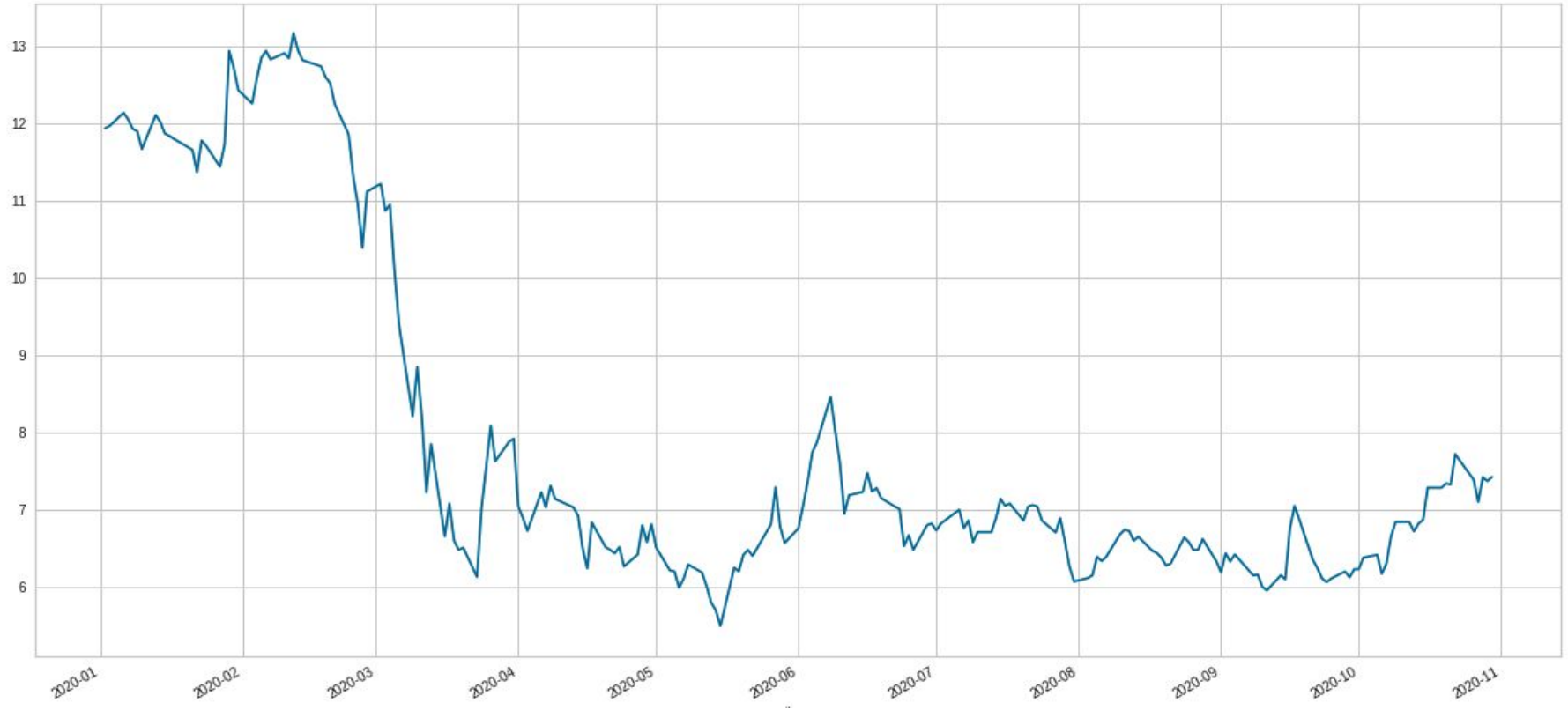
label	volume	bop	obv	mfi	quality
0	-0.359922	-0.674124	0.686409	-0.687097	0.699382
1	-0.095954	0.966205	0.785295	0.790109	0.961390
2	0.110835	0.653703	-0.975947	0.585906	-0.908150
3	3.794578	0.177968	-0.102655	0.584860	-0.509547
4	0.068192	-0.804231	-1.001559	-0.767434	-1.038356
5	-0.122527	-0.845820	0.375514	0.908284	-1.378589
6	-0.221515	0.979883	-0.187747	-0.931703	1.723838

label	volume	bop	obv	mfi
0	-0.408235	-0.722897	0.699064	-0.695031
1	0.056945	-0.791917	-1.018578	-0.846814
2	0.088682	0.640636	-0.991869	0.505555
3	3.774615	0.180592	-0.001625	0.598473
4	-0.143539	0.945236	0.813563	0.762228
5	-0.274794	0.978215	-0.075314	-0.954242
6	-0.162765	-0.849952	0.440876	0.871915

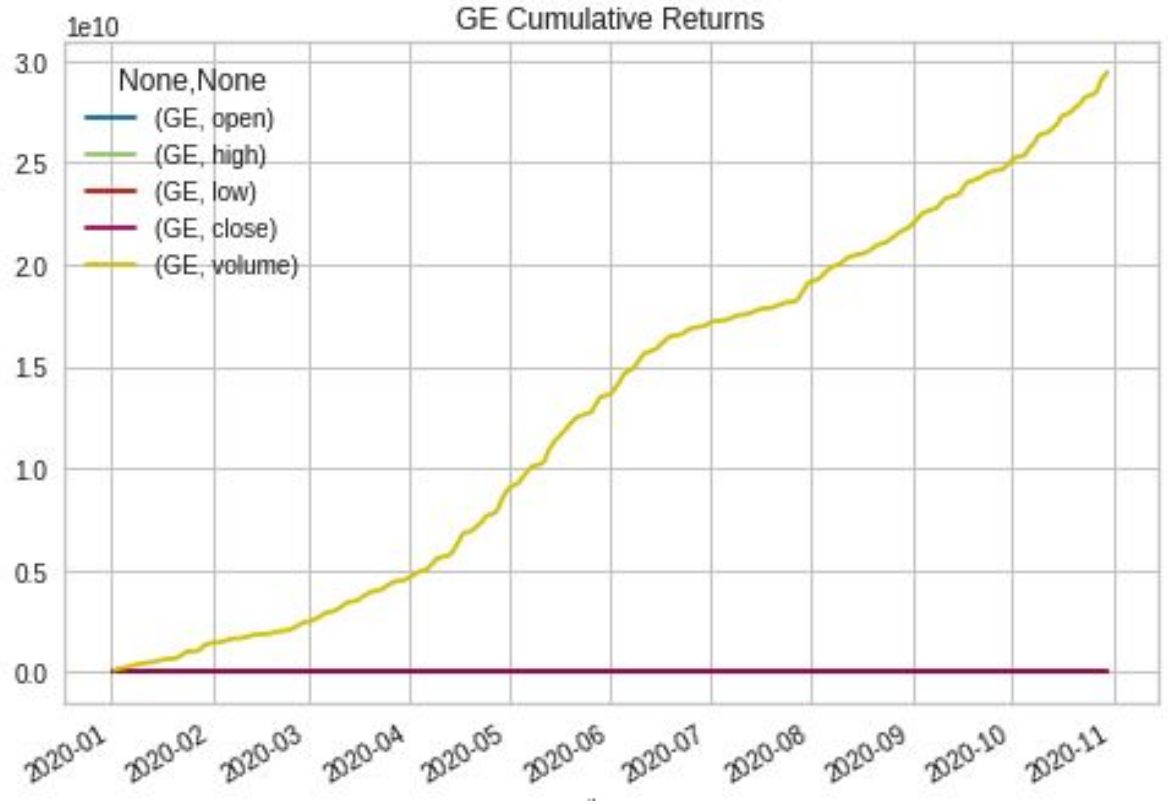
Clusters and metrics:

MFI - the lower the better(rsi)
BOP - below 0 bears/+0 bulls
OBV - trend following, should be positive

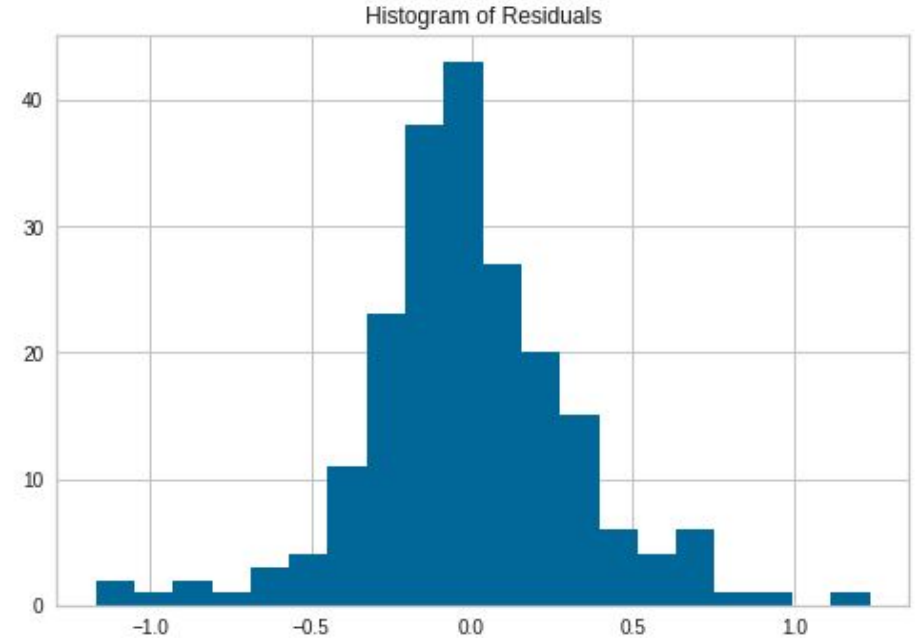
TIME SERIES ANALYSIS FOR GENERAL ELECTRICS

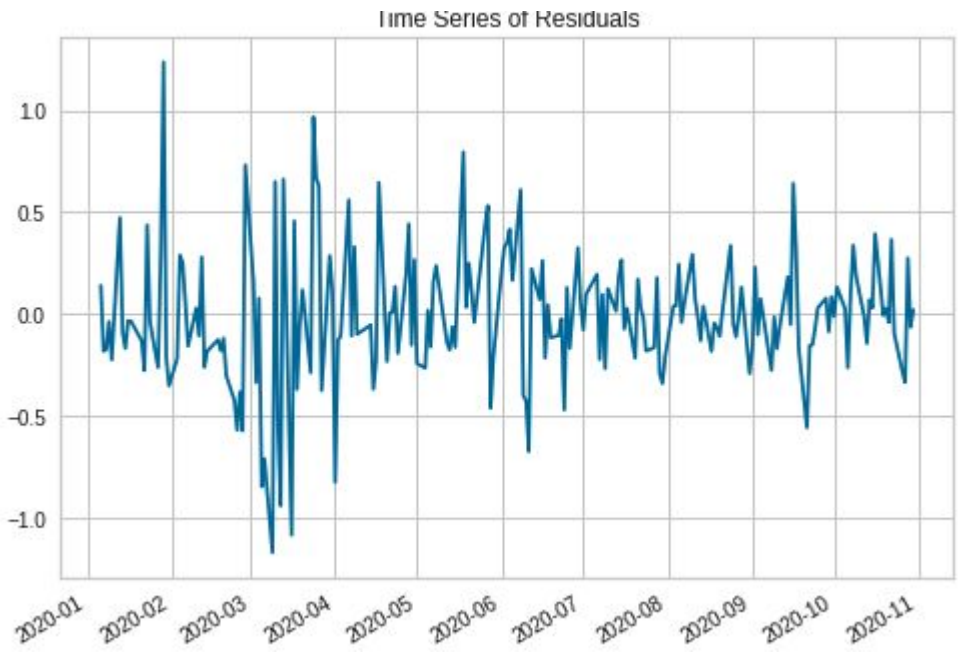


VOLUME IS UP SINCE MARCH



Residuals are random and normally distributed.
They're the difference b/n the prediction and observed quantity





ARIMA MODEL

ARIMA Model Results

Dep. Variable:	D.close	No. Observations:	210
Model:	ARIMA(0, 1, 0)	Log Likelihood	-61.645
Method:	css	S.D. of innovations	0.325
Date:	Mon, 30 Nov 2020	AIC	127.290
Time:	09:01:25	BIC	133.984
Sample:	1	HQIC	129.996

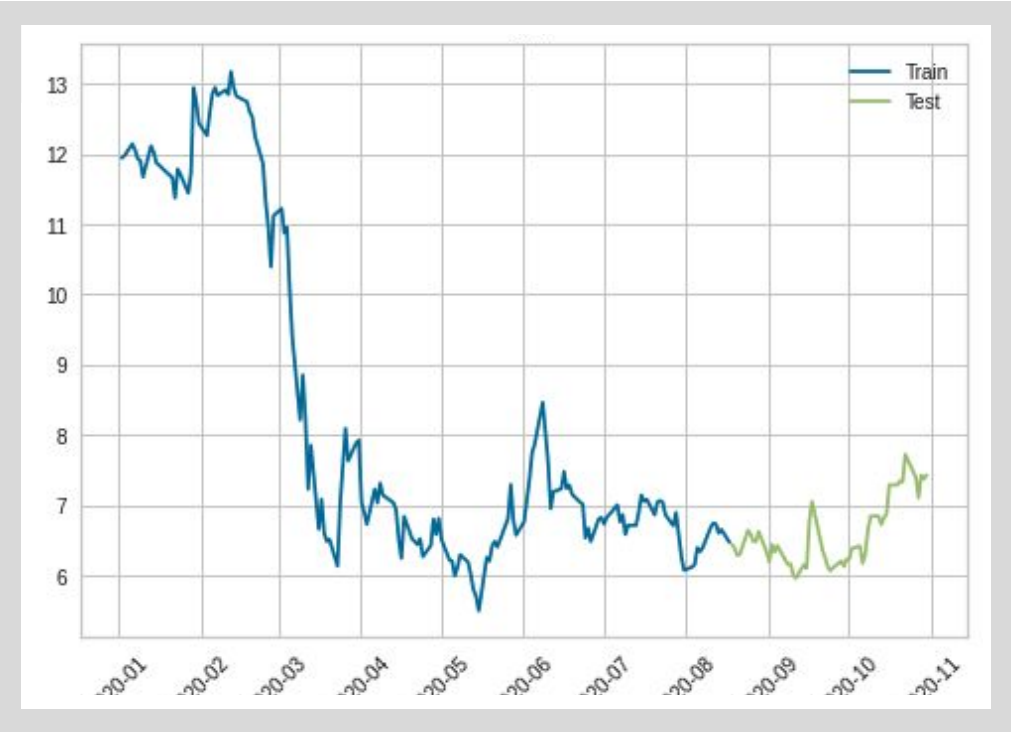
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0215	0.022	-0.960	0.337	-0.065	0.022

Residuals Description

count	2.100000e+02
mean	1.108854e-10
std	3.253008e-01
min	-1.172900e+00
25%	-1.660005e-01
50%	-1.600048e-02
75%	1.689995e-01
max	1.231500e+00

dtype: float64

The prediction



Questions?

thanks:)